From few to many: generative models for recogniti under variable pose and illumination

Georghiades, A.S. Belhumeur, P.N. Kriegman, D.J.
Dept. of Electr. Eng. & Comput. Sci., Yale Univ., New Haven, CT, USA;

This paper appears in: **Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on**

Meeting Date: 03/28/2000 - 03/30/2000
Publication Date: 28-30 March 2000
Location: Grenoble France
On page(s): 277 - 284
Reference Cited: 19
Number of Pages: xiv+560
Inspec Accession Number: 6577289

BEST AVAIL

Abstract:
Image variability due to changes in pose and illumination can seriously impair recognition. This paper presents appearance-based methods which, unlike pre appearance-based approaches, require only a small set of training images to rich representation that models this variability. Specifically, from as few as th of an object in fixed pose seen under slightly varying but unknown lighting, a an albedo map are reconstructed. These are then used to generate synthetic large variations in pose and illumination and thus build a representation usefu recognition. Our methods have been tested within the domain of face recogni subset of the Yale Face Database B containing 4050 images of 10 faces seen variable pose and illumination. This database was specifically gathered for tes generative methods. Their performance is shown to exceed that of popular ex methods

Index Terms:
albedo face recognition image reconstruction image representation learning (artifici intelligence) lighting object recognition Yale Face Database B albedo map appear methods face recognition generative models image variability object recognition p rich representation surface reconstruction synthetic images training image set varia illumination variable pose

BEST AVAILABLE COPY

Documents that cite this document

From Few to Many: Generative Models for Recognition Under Variable Pose and Illumination*

Athinodoros S. Georghiades   Peter N. Belhumeur
Departments of Electrical Engineering
and Computer Science
Yale University
New Haven, CT 06520-8267

David J. Kriegman
Beckman Institute
University of Illinois, Urbana-Champaign
Urbana, IL 61801

Abstract

Image variability due to changes in pose and illumination can seriously impair object recognition. This paper presents appearance-based methods which, unlike previous appearance-based approaches, require only a small set of training images to generate a rich representation that models this variability. Specifically, from as few as three images of an object in fixed pose seen under slightly varying but unknown lighting, a surface and an albedo map are reconstructed. These are then used to generate synthetic images with large variations in pose and illumination and thus build a representation useful for object recognition. Our methods have been tested within the domain of face recognition on a subset of the Yale Face Database B containing 4050 images of 10 faces seen under variable pose and illumination. This database was specifically gathered for testing these generative methods. Their performance is shown to exceed that of popular existing methods.

1 Introduction

An object can appear strikingly different due to changes in pose and illumination (see Figure 1). To handle this image variability, object recognition systems usually use one of the following approaches: (a) control viewing conditions, (b) employ a representation that is invariant to the viewing conditions, or (c) directly model this variability. For example, there is a long tradition of performing edge detection at an early stage since the presence of an edge at an image location is thought to be largely independent of lighting. It has been observed, however, that methods for face recognition based on finding local image features and using their geometric relation are generally ineffective [4].

Here, we consider issues in modeling the effects of both pose and illumination variability rather than trying to achieve invariance to these viewing conditions. We show how these models can be exploited for reconstructing the 3-D geometry of objects and used to significantly increase the performance of appearance-based recognition systems. We demonstrate the use of these models within the context of face recognition, but believe that they have much broader applicability.

Methods have recently been introduced which use low-dimensional representations of images of objects to perform recognition, see for example [8, 13, 19]. These methods, often termed appearance-based methods, differ from feature-based methods in that their low-dimensional representation is, in a least-squares sense, faithful to the original image. Systems such as SLAM [13] and Eigenfaces [19] have demonstrated the power of appearance-based methods both in ease of implementation and in accuracy.

Yet, these methods suffer from an important drawback: recognition of an object under a particular pose and lighting can be performed reliably *provided the object has been previously seen under similar circumstances*. In other words, these methods in their original form have no way of extrapolating to novel viewing conditions. Here, we consider the construction of a generative appearance model and demonstrate its usefulness for image-based rendering and recognition.

The presented approach is, in spirit, an appearance-based method for recognizing objects under large variations in pose and illumination. However, it differs substantially from previous methods in that it uses as few as three images of each object seen in fixed pose and under small but unknown changes in lighting. From these images, it generates a rich representation that models the object's image variability due to pose and illumination. One might think that pose variation is harder to handle because of occlusion or appearance of surface points and the non-linear warping of the image coordinates. Yet, as demonstrated by favorable recognition results, our approach can successfully generalize the concept of the illumination cone which models all the images of a Lambertian object in fixed pose under all variation in illumination [1].

New recognition algorithms based on these generative models have been tested on a subset of the Yale Face Database B (see Figure 1) which was specifically gathered for this purpose. This subset contained 4050 images of 10 faces each seen under 45 illumination conditions over nine poses. As we will see, these new algorithms outperform popular existing techniques.

*P. N. Belhumeur and A. S. Georghiades were supported by a Presidential Early Career Award, a NSF Career Award IRI-9703134, and an ARO grant DAAH04-95-1-0494. D. J. Kriegman was supported by NSF under NYI, IRI-9257990, and by an ARO grant DAAG55-98-1-0168.

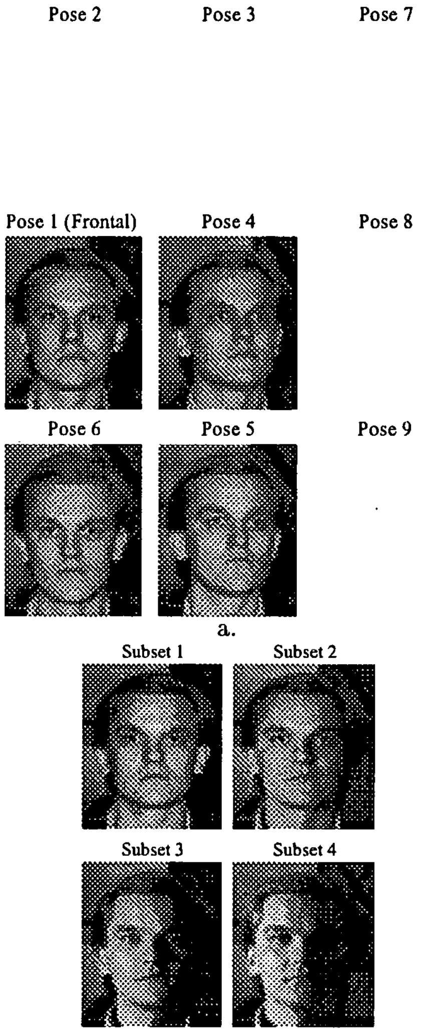


Figure 1: Example images from the Yale Face Database B, showing the variability due to pose and illumination in the images of a single individual. a. An image from each of the nine different poses; b. A representative image from each illumination subset—Subset 1 (12°), Subset 2 (25°), Subset 3 (50°), Subset 4 (77°).

2 Modeling Illumination and Pose

2.1 The Illumination Cone

In earlier work, it was shown that for a convex object with Lambertian reflectance, the set of all n-pixel images under an arbitrary combination of point light sources forms a convex polyhedral cone in the image space $\mathbb{R}^n$. This cone can be built from as few as three images [1]. Here, we outline the relevant results.

Let $\mathbf{x} \in \mathbb{R}^n$ denote an image with n pixels of a convex object with a Lambertian reflectance function illuminated by a single point source at infinity. Let $B \in \mathbb{R}^{n \times 3}$ be a matrix where each row in B is the product of the albedo with the inward pointing unit normal for a point on the surface projecting to a particular pixel in the image. A point light source at infinity can be represented by $\mathbf{s} \in \mathbb{R}^3$ signifying the product of the light source intensity with a unit vector in the direction of the light source. A convex Lambertian surface with normals and albedo given by B, illuminated by $\mathbf{s}$, produces an image $\mathbf{x}$ given by

$$\mathbf{x} = \max(B\mathbf{s}, \mathbf{0}), \qquad (1)$$

where $\max(B\mathbf{s}, \mathbf{0})$ sets to zero all negative components of the vector $B\mathbf{s}$. The pixels set to zero correspond to the surface points lying in an attached shadow. Convexity of the object's shape is assumed at this point to avoid cast shadows. Note that when no part of the surface is shadowed, $\mathbf{x}$ lies in the 3-D subspace $\mathcal{L}$ given by the span of the columns of B [8, 14, 16].

If an object is illuminated by k light sources at infinity, then the image is given by the superposition of the images which would have been produced by the individual light sources, i.e.,

$$\mathbf{x} = \sum_{i=1}^{k} \max(B\mathbf{s}_i, \mathbf{0}) \qquad (2)$$

where $\mathbf{s}_i$ is a single light source. Due to this superposition, it follows that the set of all possible images C of a convex Lambertian surface created by varying the direction and strength of an arbitrary number of point light sources at infinity is a convex cone. It is also evident from Equation 2 that this convex cone is completely described by matrix B.

Furthermore, any image in the illumination cone C (including the boundary) can be determined as a convex combination of *extreme rays* (images) given by

$$\mathbf{x}_{ij} = \max(B\mathbf{s}_{ij}, \mathbf{0}), \qquad (3)$$

where

$$\mathbf{s}_{ij} = \mathbf{b}_i \times \mathbf{b}_j. \qquad (4)$$

The vectors $\mathbf{b}_i$ and $\mathbf{b}_j$ are the rows of B with $i \neq j$. It is clear that there are at most $m(m-1)$ extreme rays for $m \leq n$ independent surface normals.

2.2 Constructing the Illumination Cone

Equations 3 and 4 suggest a way to construct the illumination cone for each object: gather three or more images in fixed pose under differing but unknown illumination without shadowing and use these images to estimate a basis for the 3-D illumination subspace $\mathcal{L}$. One way of estimation is to normalize the images to be of unit length, and then use singular value decomposition (SVD) to calculate in a least-squares sense the best 3-D orthogonal basis in the form of matrix B^*. Note that even if the columns of B^* exactly span the subspace $\mathcal{L}$, they differ from those of B by an unknown linear transformation, i.e., $B = B^*A$ where $A \in GL(3)$; for any light source, $\mathbf{x} = B\mathbf{s} = (B^*A)(A^{-1}\mathbf{s})$ [10]. Nonetheless, both B^* and B define the same illumination cone C and represent valid illumination models.

From B^*, the extreme rays defining the illumination cone C can be computed using Equations 3 and 4.

Unfortunately, using SVD in the above procedure leads to an inaccurate estimate of B^*. For even a convex object whose occluding contour is visible, there is only one light source direction (the viewing direction) for which no point on the surface is in shadow. For any other light source direction, shadows will be present. If the object is non-convex, such as a face, then shadowing in the modeling images is likely to be more pronounced. When SVD is used to find B^* from images with shadows, these systematic errors bias its estimate significantly. Therefore, an alternative way is needed to find B^* that takes into account the fact that some data values are invalid and should not be used in the estimation. For the purpose of this estimation, any invalid data can be treated as missing measurements.

The technique we use here is a combination of two algorithms. A variation of [17] (see also [11, 18]) which finds a basis for the 3-D linear subspace $\mathcal{L}$ from image data with missing elements is used together with the method in [6] which enforces integrability in shape from shading. We have modified the latter method to guarantee integrability in the estimates of the basis vectors of subspace $\mathcal{L}$ from multiple images. By enforcing integrability a surface context is introduced. Namely, the vector field induced by the basis vectors is guaranteed to be a gradient field that corresponds to a surface.

Furthermore, enforcing integrability inherently leads to more accurate estimates because there are fewer parameters (or degrees of freedom) to determine. It also resolves six out of the nine parameters of $A \in GL(3)$. The other three correspond to the generalized bas-relief (GBR) transformation parameters which cannot be resolved with illumination information alone (i.e. shading and shadows) [2]. This means we cannot recover the true matrix B and its corresponding surface, $z(x, y)$. We can only find their GBR versions $\bar{B}$ and $\bar{z}(x, y)$.

Our estimation algorithm is iterative and to enforce integrability, the possibly non-integrable vector field induced by the current estimate of B^* is, in each iteration, projected down to the space of integrable vector fields, or gradient fields [6]. To begin, let us expand the surface $\bar{z}(x, y)$ using basis surfaces (functions):

$$\bar{z}(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi(x, y; \mathbf{w}) \tag{5}$$

where $\mathbf{w} = (w_x, w_y)$ is a two dimensional index, and $\{\phi(x, y; \mathbf{w})\}$ is a finite set of basis functions which are not necessarily orthogonal. We chose the discrete cosine basis so that $\{\bar{c}(\mathbf{w})\}$ is exactly the set of the 2-D discrete cosine transform (DCT) coefficients of $\bar{z}(x, y)$.

Note that the partial derivatives of $\bar{z}(x, y)$ can also be expressed in terms of this expansion, giving

$$\bar{z}_x(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi_x(x, y; \mathbf{w}) \tag{6}$$

and

$$\bar{z}_y(x, y; \bar{c}(\mathbf{w})) = \sum \bar{c}(\mathbf{w})\phi_y(x, y; \mathbf{w}). \tag{7}$$

Since the partial derivatives of the basis functions, $\phi_x(x, y; \mathbf{w})$ and $\phi_y(x, y; \mathbf{w})$, are integrable and the expansions of $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$ share the same coefficients $\bar{c}(\mathbf{w})$, it is easy to see that $\bar{z}_{xy}(x, y) = \bar{z}_{yx}(x, y)$.

Suppose, now, we have the possibly non-integrable estimate B^* from which we can easily deduce the possibly non-integrable partial derivatives $z^*_x(x, y)$ and $z^*_y(x, y)$. These can also be expressed as a series, giving

$$z^*_x(x, y; c^*_1(\mathbf{w})) = \sum c^*_1(\mathbf{w})\phi_x(x, y; \mathbf{w}) \tag{8}$$

and

$$z^*_y(x, y; c^*_2(\mathbf{w})) = \sum c^*_2(\mathbf{w})\phi_y(x, y; \mathbf{w}). \tag{9}$$

Note that in general $c^*_1(\mathbf{w}) \neq c^*_2(\mathbf{w})$ which implies that $z^*_{xy}(x, y) \neq z^*_{yx}(x, y)$.

Let us assume that $z^*_x(x, y)$ and $z^*_y(x, y)$ are known from an estimate of B^* and we would like to find $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$ (a set of integrable partial derivatives) which are as close as possible to $z^*_x(x, y)$ and $z^*_y(x, y)$, respectively, in a least-squares sense. The goal is to minimize the following,

$$\min_{\bar{c}} \sum_{x,y} \quad (\bar{z}_x(x, y; \bar{c}) - z^*_x(x, y; c^*_1))^2 + \\ (\bar{z}_y(x, y; \bar{c}) - z^*_y(x, y; c^*_2))^2. \tag{10}$$

In other words, take a set of possibly non-integrable partial derivatives, $z^*_x(x, y)$ and $z^*_y(x, y)$, and "enforce" integrability by finding the least-squares fit of integrable partial derivatives $\bar{z}_x(x, y)$ and $\bar{z}_y(x, y)$. Notice that to get the GBR transformed surface $\bar{z}(x, y)$ we need only perform the inverse 2-D DCT on the coefficients $\bar{c}(\mathbf{w})$.

The above procedure is incorporated into the following algorithm. To begin, define the data matrix for k images of an individual to be $X = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$. If there were no shadowing, X would be rank 3 [15] (assuming no image noise), and we could use SVD to factorize X into $X = B^*S$ where S is a $3 \times k$ matrix whose columns $\mathbf{s}_i$ are the light source directions scaled by their corresponding source intensities for all k images.

Since the images have shadows (both cast and attached), and possibly saturations, we first have to determine which data values do not satisfy the Lambertian assumption. Unlike saturations, which can be simply determined, finding shadows is more involved. In our implementation, a pixel is assigned to be in shadow if its value divided by its corresponding albedo is below a threshold. As an initial estimate of the albedo we use the average of the modeling (or training) images. A conservative threshold is then chosen to determine shadows making it almost certain no invalid data is included in the estimation process, at the small expense of throwing away a few valid measurements. After finding the invalid data, the following estimation method is used:

1. Use the average of the modeling (or training) images as an initial estimate of the albedo.

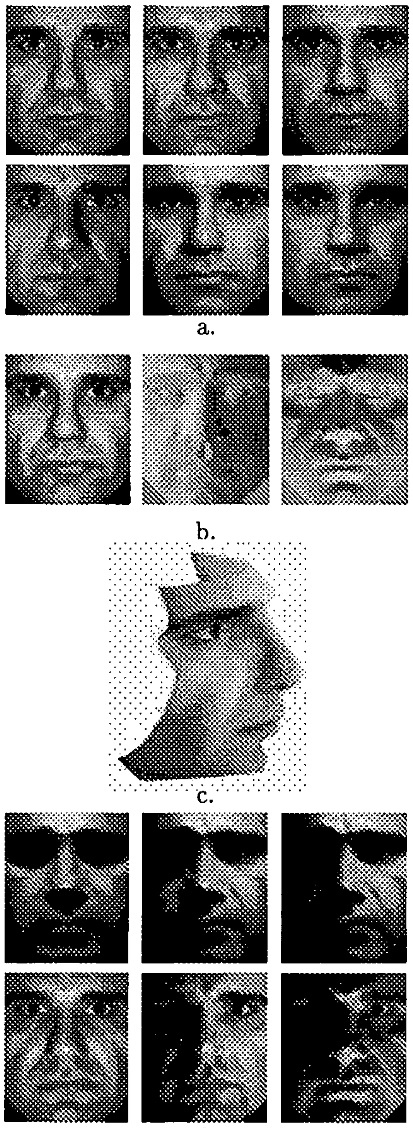


Figure 2: The process of constructing the cone C. a. The training images; b. Images corresponding to columns of $\bar{B}$; c. Reconstruction up to a GBR transformation; d. Sample images from the illumination cone under novel lighting conditions in fixed pose.

2. Without doing any row or column permutations sift out all the full rows (with no invalid data) of matrix X to form a full sub-matrix $\tilde{X}$.
3. Perform SVD on $\tilde{X}$ to get an initial estimate of S.
4. Fix S and the albedo, and estimate a possibly non-integrable set $z^*_x(x,y)$ and $z^*_y(x,y)$ using least-squares.
5. By minimizing the cost functional in Equation 10, estimate (as functions of $\bar{c}(\mathbf{w})$) a set of integrable partial derivatives $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$.
6. Fix S and use $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$ to update the albedo using least-squares.
7. Use the newly calculated albedo and the partial derivatives $\bar{z}_x(x,y)$ and $\bar{z}_y(x,y)$ to construct $\bar{B}$.
8. Then, fix $\bar{B}$ and update each of the light source directions $\mathbf{s}_i$ independently using least-squares.
9. Repeat steps 4-8 until the estimates converge.
10. Perform inverse DCT on the coefficients $\bar{c}(\mathbf{w})$ to get the GBR surface $\bar{z}(x,y)$.

In our experiments, the algorithm is well behaved, provided the input data is well conditioned, and converges within 10-15 iterations.

Figure 2 demonstrates the process for constructing the illumination cone: Figure 2.a shows six of the 19 single light source images of a face used in the estimation of matrix $\bar{B}$. Note that the light source in each image moves only by a small amount ($\pm 15^o$ in either direction) about the viewing axis. Despite this, the images do exhibit some shadowing, e.g. left and right of the nose. Figure 2.b shows the basis images of the estimated matrix $\bar{B}$. These basis images encode not only the albedo (reflectance) of the face but also its surface normal field. They can be used to construct images of the face under arbitrary and quite extreme illumination conditions. Figure 2.c shows the reconstructed surface of the face $\bar{z}(x,y)$ up to a GBR transformation. The first basis image of matrix $\bar{B}$ shown in Figure 2.b has been texture-mapped on the surface.

Figure 2.d shows images of the face generated using the image formation model in Equation 1 which has been extended to account for cast shadows. To determine cast shadows, we employ ray-tracing that uses the reconstructed GBR surface of the face $\bar{z}(x,y)$. With this extended image formation model, the generated images exhibit realistic shading and, unlike the images in Figure 2.a, have strong attached and cast shadows.

2.3 Image Synthesis Under Differing Pose and Lighting

The reconstructed surface and the illumination cones can be combined to synthesize novel images of an object under differing pose and lighting. However, one complication arises because of the generalized bas-relief (GBR) ambiguity. Even though shadows are preserved under GBR transformations [2], without resolution of this ambiguity, images with non-frontal view-point synthesized from a GBR reconstruction will differ from a valid image by an affine warp of image coordinates. (It is affine because GBR is a 3-D affine transformation and the weak perspective imaging model assumed here is linear.) Since the affine warp is an image transformation, one could perform recognition over variation in viewing direction and affine image transformations. Alternatively, one can attempt to resolve the GBR ambiguity to obtain a Euclidean reconstruction using class-specific information. In our experiments with faces, we essentially try to fit the GBR reconstructions to a canonical face. We take advantage of the left-to-right symmetry of faces and the fairly constant ratios

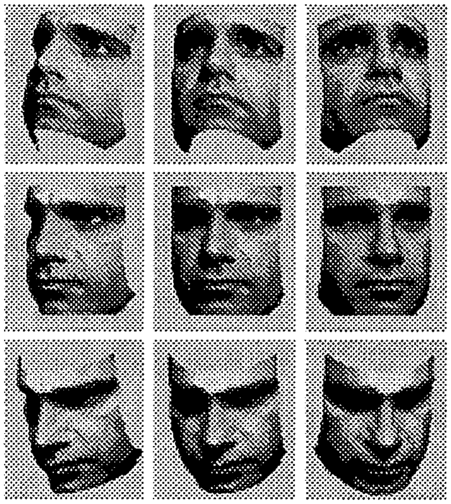


Figure 3: Synthesized images under variable pose and lighting. The representation was constructed from the images in Figure 2.a.

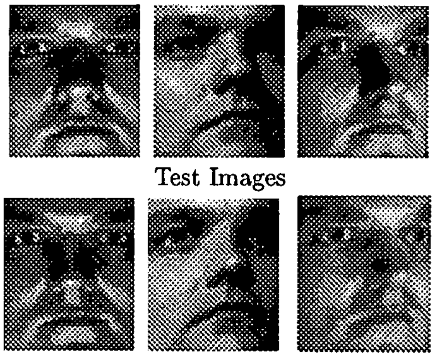
of distances between facial features such as the eyes, the nose, and the forehead to resolve the three parameters of the GBR ambiguity. Once resolved, it is a simple matter to use ray-tracing techniques to render synthetic images under variable pose and lighting.

Figure 3 shows synthetic images of the face under novel pose and lighting. These images were generated from the images in Fig. 2.a where the pose is fixed and there are only small, unknown variations in illumination. In contrast, the synthetic images exhibit not only large variations in pose but also a wide range in shading and shadowing.

3 Representations for Recognition

It is clear that for every pose of the object, the set of images under all lighting conditions is a convex cone. Therefore, the previous section provides a natural way for generating synthetic representations of objects suitable for recognition under variable pose and illumination. For every sample pose of the object, generate its illumination cone and with the union of all the cones form its representation.

However, the number of independent normals in B can be large (more than a thousand) hence the number of extreme rays needed to completely define the illumination cone can run in the millions (see Section 2). Therefore, we must approximate the cone in some fashion; in this work, we choose to use a small number of extreme rays (images). The hope is that a subsampled cone will provide an approximation that negligibly decreases recognition performance; in our experience, around 80 images are sufficient, provided that the corresponding light source directions $\mathbf{s}_{ij}$ are more or less uniform on the illumination sphere. The resulting cone C^* is a subset of the object's true cone C for a particular pose.



Test Images

Closest image in cone representation

Figure 4: TOP ROW: Three images from the test set. BOTTOM ROW: The closest reconstructed image from the representation. Note that these images are not explicitly stored, but lie within the closest matching linear subspace.

Another simplifying factor that can reduce the size of the representation is the assumption of a weak perspective imaging model. Under this model, the effect of pose variation can be decoupled into that due to image plane translation, rotation, and scaling (a similarity transformation), and that due to the viewpoint direction. Within a face recognition system, the face detection process generally provides estimates for the image plane transformations. Neglecting the effects of occlusion or appearance of surface points, the variation due to viewpoint can be seen as a non-linear warp of the image coordinates with only two degrees of freedom.

Yet, recognition using this representation consisting of sub-sampled illumination cones will still be costly since computing distance to a cone is $O(n\,e^2)$, where n is the number of pixels and e is the number of extreme rays (images). From an empirical study, it was conjectured in [1] that the cone for typical objects is flat (i.e., all points lie near a low-dimensional linear subspace), and this was confirmed for faces in [5]. Hence, an alternative is to model a face in fixed pose but over all lighting conditions by a low-dimensional linear subspace. Finally, for a set of sample viewing directions, we construct subspaces which approximate the corresponding cones. We chose to use an 11-D linear subspace for each pose since 11 dimensions capture over 99% of the variation in the sample extreme rays. Recognition of a test image $\mathbf{x}$ is then performed by finding the closest linear subspace to $\mathbf{x}$. Figure 4 shows the closest match for images of an individual in three poses. This figure qualitatively demonstrates how well the union of 11-D subspaces approximates the true cones.

For the experimental results reported below, subspaces were constructed by sampling the viewing sphere at 4° intervals over the elevation from $-24°$ to $+24°$ and the azimuth from $-4°$ to $+28°$ about frontal. As a final speed-up, the 117 11-D linear subspaces were projected down to a 100-dimensional subspace of the image space whose basis vectors were computed using

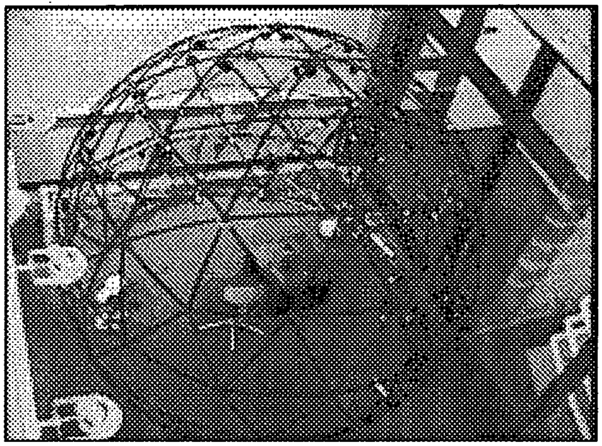


Figure 5: A geodesic dome with 64 strobes used to gather images under variable illumination and pose.

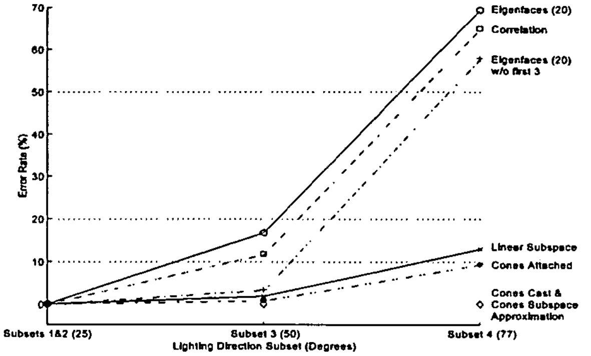
SVD. In summary, each person's face was represented by the union of 117 11-D linear subspaces within a 100-dimensional subspace of the image space. Recognition was then performed by computing the distance of a test image to each 100-D subspace plus the distance to the 11-D subspaces within the 100-D space.

4 Recognition Results

The experimentation reported here was performed on the Yale Face Database B. For capturing this database, we have constructed a geodesic lighting rig with 64 computer controlled xenon strobes shown in Figure 5. With this rig, we can modify the illumination at frame rates and capture images under variable pose and illumination. Images of ten individuals were acquired under 64 different lighting conditions in nine poses (frontal pose, five poses at 12° and three poses at 24° from the camera's axis). Of the 64 images per person in each pose, 45 were used in our experiments, a total of 4050 images. The images from each pose were divided into 4 subsets (12°, 25°, 50° and 77°) according to the angle of the light source with the camera's axis (see Figure 1). Subset 1 (respectively 2, 3, 4) contains 70 (respectively 120, 120, 140) images per pose. Throughout, the 19 images of Subsets 1 and 2 from the frontal pose of each face were used as training images for generating its representation.

4.1 Extrapolation in Illumination

The first set of experiments was performed under fixed pose on the 450 images from the frontal pose (45 per person). This was to compare three other recognition methods to the illumination cones representation. From a set of face images labeled with the person's identity (*the learning set*) and an unlabeled set of face images from the same group of people (*the test set*), each algorithm is used to identify the person in the test images. For more details about the comparison algorithms, see [3] and [7]. We assume that each face has been located and aligned within the image.



EXTRAPOLATION IN ILLUMINATION

Method	Error Rate (%) vs. Illumination		
	Subsets 1 & 2	Subset 3	Subset 4
Correlation	0.0	11.7	65.0
Eigenfaces	0.0	16.7	69.3
Eigenfaces w/o 1st 3	0.0	3.3	57.9
Linear subspace	0.0	1.7	12.9
Cones-attached	0.0	0.8	9.3
Cones-cast (Subspace Approx.)	0.0	0.0	0.0
Cones-cast	0.0	0.0	0.0

Figure 6: **Extrapolation in Illumination:** Each of the methods is trained on images with near frontal illumination (Subsets 1 and 2) from Pose 1 (frontal pose). This graph shows the error rates under more extreme light source conditions in fixed pose.

The simplest recognition scheme is a nearest neighbor classifier in the image space [4]. An image in the test set is recognized (classified) by assigning to it the label of the closest point in the learning set, where distances are measured in the image space. When all of the images are normalized to have zero mean and unit variance, this procedure is also known as Correlation.

A technique now commonly used in computer vision—particularly in face recognition—is principal components analysis (PCA) which is popularly known as *Eigenfaces* [8, 12, 13, 19]. One proposed method for handling illumination variation in PCA is to discard the three most significant principal components; in practice, this yields better recognition performance [3]. For both the Eigenfaces and Correlation tests, the images were normalized to have zero mean and unit variance, as this improved the performance of these methods. This also made their results independent of light source intensity. For the Eigenfaces method, we used 20 principal components; recall that performance approaches correlation as the dimension of the feature space is increased [3, 13]. Error rates are also presented when the principal components four through twenty-three were used.

A third approach is to model the illumination variation of each face with the three-dimensional linear subspace $\mathcal{L}$ described in Section 2.1. To perform recognition, we simply compute the distance of the test image to each linear subspace and choose the face corresponding to the shortest distance. We call this recognition scheme the *Linear Subspace* method [2]; it is a variant of the photometric alignment method proposed in [16] and is related to [9, 14]. While this models the variation in image intensities when the surface is completely illuminated, it does not model shadowing.

Finally, recognition is performed using the illumination cone representation. In fact, we tested on three variations. In the first (Cones-attached), the representation was constructed without cast shadows, so the extreme rays are generated directly from Equation 3. In the second variation (Cones-cast), the representation was constructed as described in Section 2.2 where we employed ray-tracing that uses the reconstructed surface of a face $\bar{z}(x, y)$ to determine cast shadows. In both variations, recognition was performed by computing the distance of the test image to each cone and choosing the face corresponding to the shortest distance. Since cones are convex, the distance can be found by solving a convex optimization problem (see [7]).

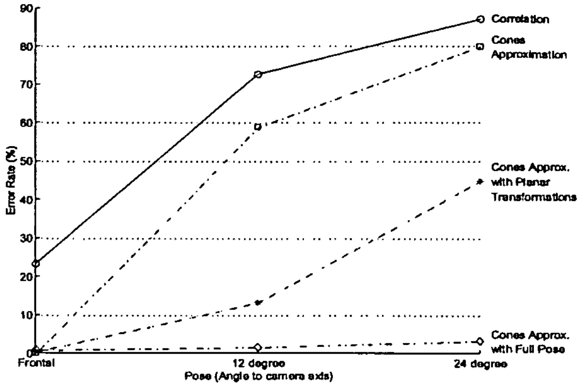
In the last variation, the illumination cone of each face with cast shadows C^* is approximated by an 11-D dimensional linear subspace (Cones-cast subspace approximation). As mentioned before, it was empirically determined that 11 dimensions capture over 99% of the variance in the sample extreme rays. The basis vectors for this space are determined by performing SVD on the extreme rays in C^* and then picking the 11 eigenvectors associated with the largest singular values. Recognition was performed by computing the distance of the test image to each linear subspace and choosing the face corresponding to the shortest distance. Using the cone subspace approximation reduces both the storage and the computational time. Since the basis vectors of each subspace are orthogonal the computational complexity is only $O(n\,m)$ where n is the number of pixels and m is the number of the basis vectors.

Similar to the extrapolation experiment described in [3], each method was trained on samples from Subsets 1 and 2 (19 samples per person) and then tested on samples from Subsets 3 and 4. Figure 6 shows the results from this experiment. (This test was also performed on the Harvard Robotics Lab face database and was reported in [7].) Note that the cone subspace approximation performed as well as the raw illumination cones without any mistakes on 450 images. This supports the use of low dimensional subspaces in the full representation of Section 3 that models image variations due to viewing direction and lighting.

4.2 Recognition Under Variable Pose and Illumination

Next, we performed recognition experiments on images in which the pose varies as well as illumination. Images from all nine poses in the database were used in these tests. Four recognition methods were compared on 4050 images. Each method was trained on images with near frontal illumination (Subsets 1 and 2) from the frontal pose, and tested on all images from all nine poses—an extrapolation in both pose and illumination.

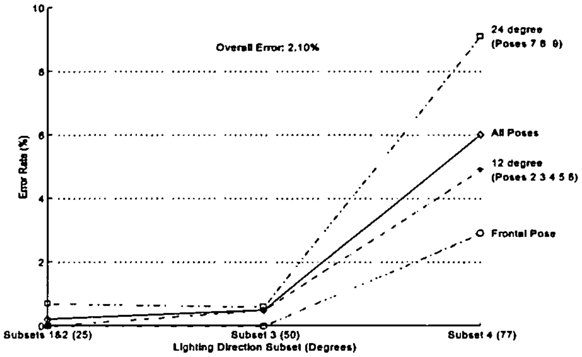
The first method was Correlation as described in the previous section. The next one (Cones approximation) modeled a face with an 11-D subspace approximation of the cone (with cast shadows) in the frontal pose. No effort was done to accommodate pose during recognition, not even a search in image plane transformations. The next method (Cones approximation with planar transformations) also modeled a face with an 11-D subspace approximation of the cone in the frontal pose, but unlike the previous method, recognition was performed over variations of planar transformations. Finally, a face was modeled with the representation described in Section 3. Each of the 10 individuals was represented by a 100-D subspace which contained 117 11-D linear subspaces each modeling the variation in illumination for each sampled view-point. As with the previous method, recognition was performed over a variation of planar transformations. The results of these experiments are shown in Figure 7. Note that each reported error rate is for *all* illumination subsets (1 through 4). Figure 8, on the other hand, shows the break-down of the results of the last method for different poses against



EXTRAPOLATION IN POSE

Method	Error Rate (%) vs. Pose		
	Frontal (Pose 1)	12° (Poses 2 3 4 5 6)	24° (Poses 7 8 9)
Correlation	23.3	72.6	87.0
Cones Approximation	0.0	58.8	79.9
Cones Approx. with Planar Transformations	0.4	13.3	44.8
Cones Approx. with Full Pose	0.9	1.6	3.3

Figure 7: **Extrapolation in Pose:** Error rates as the viewing direction becomes more extreme. Again, the methods were trained on images with near frontal illumination (Subsets 1 and 2) from Pose 1 (frontal pose). Note that each reported error rate is for *all* illumination subsets (1 through 4).



Error Rates (%)			
	Lighting Variation		
Pose	Subsets 1 & 2	Subset 3	Subset 4
Frontal (Pose 1)	0.0	0.0	2.9
12° (Poses 2 3 4 5 6)	0.0	0.5	4.9
24° (Poses 7 8 9)	0.7	0.6	9.1
All Poses	0.2	0.5	6.0

Figure 8: Error rates for different poses against variable lighting using the representation of Section 3.

variable illumination. As demonstrated in Figure 7, the method of cone subspace approximation with planar transformations performs reasonably well for poses up to 12° from the viewing axis but fails when the viewpoint becomes more extreme.

We note that in the last two methods the search in planar transformations did not include image rotations (only translations and scale) to reduce computational time. We believe that the results would improve if image rotations were included or even if the view-point space and illumination cones were more densely sampled and the 11-D subspaces were not projected down to a 100-D subspace.

5 Discussion

In constructing the representation of an object from a small set of training images, we have assumed that the object's surface exhibited a Lambertian reflectance function. Although our results support this assumption, more complex reflectance functions may yield better recognition results. Other exciting domains for these representations include facial expression recognition and object recognition with occlusions.

References

[1] P. Belhumeur and D. Kriegman. What is the set of images of an object under all possible illumination conditions. *Int. J. Computer Vision*, 28(3):245–260, July 1998.

[2] P. Belhumeur, D. Kriegman, and A. Yuille. The bas-relief ambiguity. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 1040 1046, 1997.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 19(7):711–720, 1997. Special Issue on Face Recognition.

[4] R. Brunelli and T. Poggio. Face recognition: Features vs templates. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 15(10):1042 1053, 1993.

[5] R. Epstein, P. Hallinan, and A. Yuille. 5+/-2 eigenimages suffice: An empirical investigation of low-dimensional lighting models. In *Physics Based Modeling Workshop in Computer Vision*, Session 4, 1995.

[6] R. T. Frankot and R. Chellapa. A method for enforcing integrabilty in shape from shading algorithms. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 10(4):439–451, 1988.

[7] A. Georghiades, D. Kriegman, and P. Belhumeur. Illumination cones for recognition under variable lighting: Faces. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 52 59, 1998.

[8] P. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 995 999, 1994.

[9] P. Hallinan. *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions*. PhD thesis, Harvard University, 1995.

[10] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *J. Opt. Soc. Am. A*, 11(11):3079–3089, Nov. 1994.

[11] D. Jacobs. Linear fitting with missing data: Applications to structure from motion and characterizing intensity images. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 1997.

[12] L. Sirovitch and M. Kirby. Low-dimensional procedure for the characterization of human faces. *J. Optical Soc. of America A*, 2:519 524, 1987.

[13] H. Murase and S. Nayar. Visual learning and recognition of 3-D objects from appearence. *Int. J. Computer Vision*, 14(5 24), 1995.

[14] S. Nayar and H. Murase. Dimensionality of illumination manifolds in appearance matching. In *Int. Workshop on Object Representations for Computer Vision*, page 165, 1996.

[15] A. Shashua. *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, MIT, 1992.

[16] A. Shashua. On photometric issues to feature-based object recognition. *Int. J. Computer Vision*, 21:99–122, 1997.

[17] H. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 17(9):854 867, September 1995.

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137 154, 1992.

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3(1):71–96, 1991.

Face recognition using statistical models

Edwards, G.J. Lanitis, A. Taylor, C.J. Cootes, T.F.
Dept. of Med. Biophys., Manchester Univ.

This paper appears in: **Image Processing for Security Applications (Dige 1997/074), IEE Colloquium on**

Meeting Date: 03/10/1997
Publication Date: 10 March 1997
Location: London UK
On page(s): 2/1 - 2/6
Reference Cited: 9
Number of Pages: 70
Inspec Accession Number: 5584685

Abstract:
We describe the use of flexible models for the representation of shape and gr appearance of human faces. The models are controlled by a small number of which can be used to code the overall appearance of a face for image compre classification. Shape and grey-level appearance are included in a single mode Discriminant analysis allows the isolation of variation important for classificati identity. We have performed both face recognition and face synthesis experin present the results in this paper

Index Terms:
classification data compression discriminant analysis face recognition face synthesi
classification image coding image compression isolation of variation statistical mod
classification data compression discriminant analysis face recognition face synthesi
classification image coding image compression isolation of variation statistical mod

Documents that cite this document
There are no citing documents available in IEEE Xplore at this time.

Copyright © 2004 IEEE — All rights reserved

FACE RECOGNITION USING STATISTICAL MODELS

G J Edwards, A Lanitis, C J Taylor and T F Cootes*

We describe the use of flexible models for the representation of shape and grey-level appearance of human faces. The models are controlled by a small number of parameters, which can be used to code the overall appearance of a face for image compression and classification. Shape and grey-level appearance are included in a single model. Discriminant analysis allows the isolation of variation important for classification of identity. We have performed both face recognition and face synthesis experiments and present the results in this paper.

Introduction

A successful face recognition system should be able to locate a face, and classify its identity, regardless of factors such as pose, lighting and expression variation. Human faces are highly variable objects, both in terms of the different appearance of individuals, and the variation present in any individual face. In this context, the analysis of human faces presents a difficult machine vision task. As a result of this difficulty, some researchers have concentrated on particularly constrained applications; contributing little to overall progress. Others have attempted to tackle the various generic problems independently; the drawback of this approach is that the effects of all the sources of variability are compounded, so it is extremely difficult to extract a description for one characteristic of interest (e.g. individual appearance) which is not sensitive to others. (e.g. facial expression, lighting and pose).[1] Many current techniques can be found in the review by Chellapa et al.[2]

Rather than trying to separate face analysis into various goals, such as feature location, person identification, expression recognition, lighting correction, etc., we have developed a unified approach. The basis for this is a compact, parameterized model of facial appearance, which accounts for all the important, systematic sources of variability. Our approach consists of both modeling, in which flexible appearance models of facial appearance are generated, and interpretation, in which the models are used to analyze information content of the face image, such as the identity of the individual.

Modeling Shape Appearance

In order to understand the appearance of faces, we model both the shape and grey-level appearance of a training set of face images. All the models used in our system are of the same mathematical form. Each of the training examples is represented by N variables:

$$\mathbf{X}_i = (x_{1i}, x_{2i}, \ldots, x_{Ni}) \qquad (1)$$

where x_{ki} is the *k*th variable in the *i*th example

When modeling the shape of faces, these variables represent the positions of key landmark points on the images in the training set. From the training examples we build a *Point Distribution Model (PDM)*[3]. Each training image is marked with a set of 144 labeled points, corresponding to specific facial features. An example of a face overlaid with landmark points is shown in Figure 1. Given a set of these training vectors, the average example, $\mathbf{X}_{mean}$ is calculated and the deviation of each example from the mean established. A principal component analysis of the covariance matrix of the deviations reveals the main linearly independent *modes of variation* of face shape, which together represent nearly all the variation in the training set (typically 30 modes for 99.5% of a training set containing 400 examples) . Any training example, $\mathbf{X}_i$ can be approximated by using:

$$\mathbf{X}_i = \mathbf{X}_{mean} + \mathbf{Pb} \qquad (2)$$

* The authors are in The Department of Medical Biophysics, University of Manchester, Oxford Rd., Manchester. Tel No. 0161 275 5130. Email {gje,lan,bim,cjt}@sv1.smb.man.ac.uk

© 1997 The Institution of Electrical Engineers.

Printed and published by the IEE, Savoy Place, London WC2R 0BL, UI

where P is a matrix of unit eigenvectors of the covariance of deviations and **b** is a vector of eigenvector weights (these are referred to as *model parameters*). By modifying **b**, new instances of the model can be generated; if the elements of **b**, are kept within a few standard deviations of the mean over the training set, then the corresponding model instances are plausible examples of the modeled objects. By varying each model parameter over this limited range, we can illustrate thc modes of variation of the model, as shown in Figure 2.

Since the columns of **P** are orthogonal, $\mathbf{P}^T\mathbf{P} = \mathbf{I}$, and equation 2 can be solved with respect to **b**:

$$\mathbf{b} = \mathbf{P}^T(\mathbf{X} - \mathbf{X}_{mean}) \qquad (3)$$

Equation 3 can be used to transform an example **X** into model parameters.

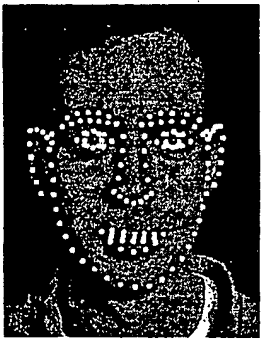


Figure 1. Training Example Overlaid with Landmark Points

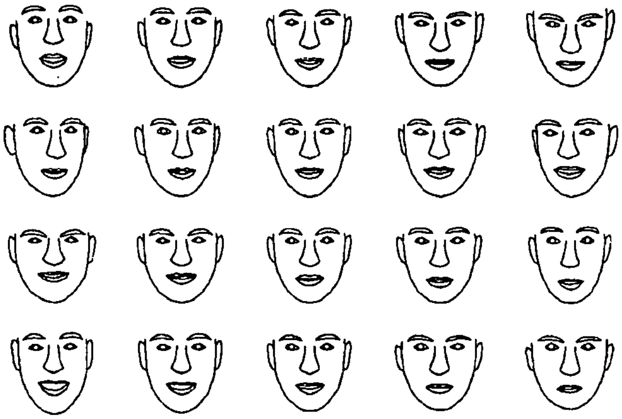


Figure 2. First 4 modes of shape variation shown varying horizontally

Modeling Full Appearance

The same statistical method can be used to model the grey-level appearance of faces. We wish to model grey-level appearance independently of shape. To do this, we first apply a warping algorithm[4] developed by Bookstein based on thin-plate splines, which warps each example to the mean shape, in such a way that grey-level changes around each landmark are kept to a minimum. Each training example is then represented by the pixel intensity values in the mean shape patch. After applying Principal Component Analysis, as for the shape model, it is possible to represent 95% of the variation in the 400 example training set by 70 parameters.

In order to complete the model of facial appearance we combine the shape and grey-level models to produce a *Combined Appearance Model.* Given the shape and grey-level models we obtain the model parameters for each training example, using Equation 3, and concatenate the two parameter vectors. A principal component analysis of these concatenated vectors over all the training examples leads to a single model describing both shape and grey-level variation. This combined model captures 95% of the variance in the 400 example training set using 55 parameters. The model fully accounts for correlation between shape and grey-level appearance.

Figure 3 shows the major modes of orthogonal variation of this combined appearance model. It can be seen that in each of the modes of variation, several sources of variation are compounded, showing variation due to pose, expression, lighting and ID.

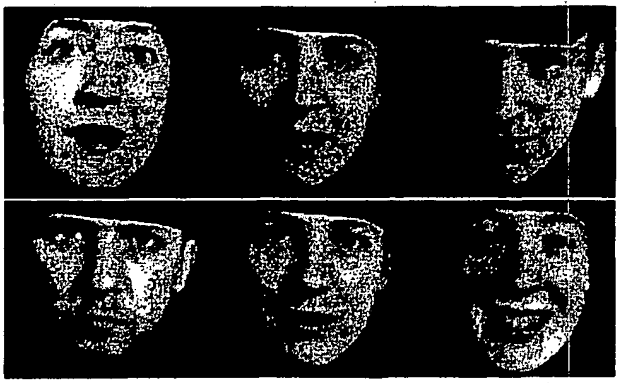


Figure 3. First 2 Modes of Full Variation Shown Horizontally +/- 3 SD's.

Locating and Tracking Faces

The flexible shape model can be used in an *Active Shape Model*(ASM) search^5 to locate and track faces in static images and image sequences. During the training phase a model is built of the expected grey-level variation around each landmark point. In order to locate and track a face, an instance of the face model is placed in the initial image and is allowed to interact until it fits the shape of the face. Each model point attempts to move towards the best local match, but the shape of the whole set of points can only be changed by varying the shape parameters, thus ensuring that resulting shapes are similar to those encountered in the training set. In order to allow a greater search range, the algorithm can be performed at lower resolution levels until a certain degree of fit is achieved, before switching to higher resolutions. Figure 4 shows some examples of face location.

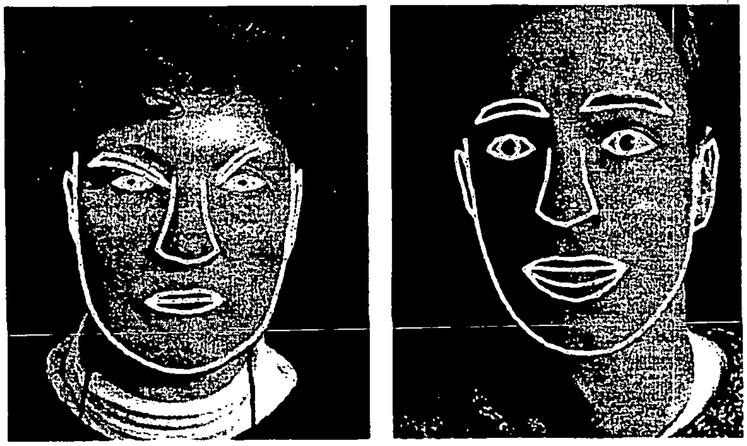


Figure 4. Examples of Successful Face Location

Identifying Faces

When a test image is presented to our system, the flexible shape model is used to locate the face and features automatically, using a *Multi-Resolution Active Shape Model* search. In the results presented in this paper the user is asked to indicate the approximate position of the nose, and the shape model is overlaid. Euclidean transformations and deformations are applied until the model is fitted to the face presented. Lanitis et al.[6] have shown that it is possible to locate the face without any user initialization using global optimization techniques. Once the model has located the face, we find the appearance model parameters and local grey-level model parameters using Equation 3; These are used for classification.

Since each of the variables shows variation due to different sources of variation, it is important to emphasize variation which is important for recognition, by using discriminant analysis[7]. This can be done by calculating the Canonical Discriminant Functions[8], or in it's simplest form, by assigning a class label based on the Mahalanobis distance. The Mahalanobis distance measure automatically assigns higher weights to those variables which showed a greater deal of inter-class (inter-person) variation during the training phase.

Lanitis et al.[9] have performed person recognition trials, the main results of which are outlined below:

The training set consisted of **30 individuals**, *10 images* of each.
The main test set consisted of *10 unseen images* of each of the **30 individuals.**
A secondary test set consisted of 3 *images* each of 'difficult' images of the **30 individuals.**

For all individuals, the test and training images showed varying pose, lighting and expression conditions. The 'difficult' test images showed partial occlusion of the face. Some examples of the images used are shown in Figure 5.

Normal Test Set (200 Images)	- Correct	**95.5%**
	- Correct within best 3	**99.0%**
Difficult Test Set (60 Images)	- Correct	**43.3%**
	- Correct within best 3	**71.6%**

These results include errors where either the face was not correctly located, or where it was correctly located but incorrectly identified.

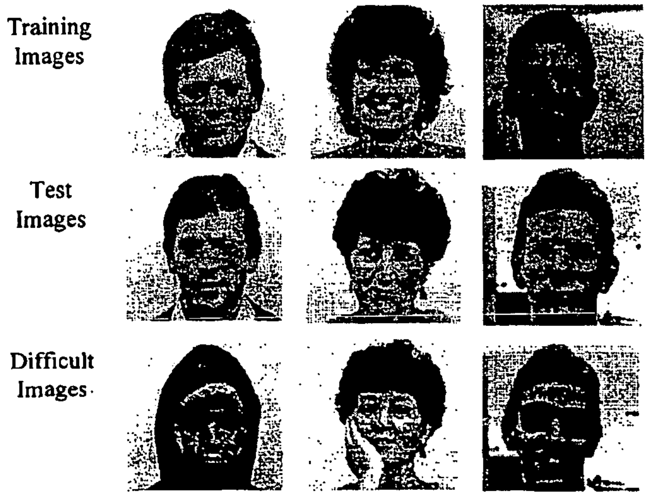


Figure 5. Training and Test Examples

Reconstruction and Coding

Given a set of combined appearance model parameters it is possible to reconstruct a face image. Figure 6 shows a face image, together with its parameterized reconstruction. The reconstruction is made from 55 parameters, the original face image is at 320x256 resolution. This represents a very high degree of compression.

We have (Edwards et al.[8]) addressed the problem of extracting specific types of variation from the combined models, in order to make the models more specific for particular applications. Using Canonical Discriminant Analysis it is possible to define a set of orthogonal modes of variation which correspond only to one type of variation, for example, change of identity. Also, having found these modes, it is possible to remove them from the model. The resulting model allows us to manipulate face images without changing their identity. In Figure 7, we show an example of a face, and some manipulations of that face, by varying parameters which have been selected so as not to change the identity of the face. Thus, given a single example of a face, we can synthesize its appearance under different conditions of pose, lighting, and expression. When only those modes of variation which do not change identity are considered, the model achieves further compression, reducing to just 21 parameters.

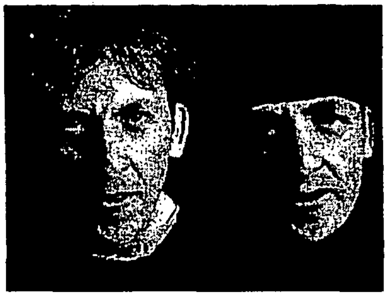


Figure 6. Face with its Parameterized Reconstruction

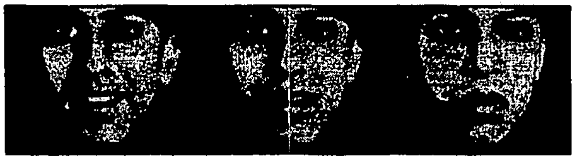


Figure 7. Face Image Manipulated by Varying Parameters.which don't Affect Identity

Conclusions

We have presented a system, which can be used for locating and tracking faces, coding, reconstruction, and identification. Our recognition results are very encouraging, especially considering the allowed variation in pose, lighting and expression. The statistical approach allows unimportant variation to be dealt with automatically. Our system copes with all aspects of face image processing within a single framework. The ability to locate and identify faces has potential for powerful security applications, particularly in access control and person monitoring. A major benefit of the system is that it is likely to be entirely passive, requiring no interaction (or even knowledge of presence) unlike, say, a keycard, or fingerprint reader.

The small number of parameters required to code a face image allows very compact storage. Access cards, Bank cards, and such like, could store encrypted appearance parameters of their owner in the magnetic strip, the reader of which would display an image of the true owner to the Bank teller or Shop assistant, virtually eliminating the use of fake or stolen credit cards. The high compression would also allow very fast comparison of faces with image databases. It is possible to envisage police and security services equipped with small CCD cameras able to instantly compare a live person with a database of faces. Statistical models are equally useful for reconstruction and manipulation as shown in Figures 6 and 7. The ability to manipulate images in a photo-realistic way has potential usefulness in forensic techniques such as photo-fit.

Acknowledgements
The authors would like to thank everyone who helped compile the face databases including Dr. Jane Whittaker, Dept. of Child Psychiatry, University of Manchester, and Mr. Ian Wilks. This work is supported in part by EPSRC and British Telecom Research Laboratories, Ipswich.

REFERENCES

[1] Y. Moses, Y. Adini and S. Ulman. "Face Recognition: The problem of compensating for changes in illumination direction". Procs of the European Conference on Computer Vision, Vol 1, pp 286-296, ed. J. Eklundh, Springer-Verlag, 1994.

[2] R. Chellapa, C.L. Wilson and S. Sirohey. "Human and Machine Recognition of Faces: A Survey. Procs of the IEEE, Vol 83, no. 5, 1995.

[3] T.F. Cootes, C.J. Taylor, D.H. Cooper and J. Graham. "Active Shape Models – Their Training and Application". Computer Vision, Graphics, and Image Understanding, Vol. 61, No. 1, pp. 38-59, 1995.

[4] F.L. Bookstein. "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 11, no 6. pp 567-585, 1989.

[5] T.F. Cootes, A. Hill, C.J. Taylor and J. Haslam. "Use of Active Shape Models for Locating Structures in Medical Images", Image and Vision Computing, 1994, Vol 12(6), pp. 355-365.

[6] A. Lanitis, "Automatic Location of Facial Characteristics using Genetic Algorithms. Wolfson Image Analysis Unit Internal Report, June 1993.

[7] B.F.J. Manly. "Multivariate Statistical Methods, A Primer." Chapman and Hall, 1986.

[8] G.J. Edwards, A. Lanitis, C.J. Taylor, T.F. Cootes, "Modelling the Variability in Face Images". Procs. International Conference on Automatic Face and Gesture Recognition, 1996, pp. 328-333, IEEE Computer Society Press.

[9] A. Lanitis, C.J. Taylor, T.F. Cootes. "Automatic Face Identification System using Flexible Appearance Models." Image and Vision Computing, Vol. 13, No. 5, pp. 393-401, 1995.

IEEE HOME | SEARCH IEEE | SHOP | WEB ACCOUNT | CONTACT IEEE

IEEE

Membership Publications/Services Standards Conferences Careers/Jobs

IEEE Xplore® RELEASE 1.6

Welcome
United States Patent and Trademark Office

Help FAQ Terms IEEE Peer Review Quick Links » ABS

Welcome to IEEE Xplore®

○- Home
○- What Can I Access?
○- Log-out

Tables of Contents

○- Journals & Magazines
○- Conference Proceedings
○- Standards

Search

○- By Author
○- Basic
○- Advanced

Member Services

○- Join IEEE
○- Establish IEEE Web Account
○- Access the IEEE Member Digital Library

Search Results [PDF FULL-TEXT 1288 KB] PREV NEXT DOWNLOAD CITATION

Request Permissions RIGHTSLINK

Illumination-based image synthesis: creating novel of human faces under differing pose and lighting

Georghiades, A.S. Belhumeur, P.N. Kriegman, D.J.
Centre for Comput. Vision & Control, Yale Univ., New Haven, CT, USA;

This paper appears in: **Multi-View Modeling and Analysis of Visual Scene (MVIEW '99) Proceedings. IEEE Workshop on**

Meeting Date: 06/26/1999
Publication Date: 26 June 1999
Location: Fort Collins, CO USA
On page(s): 47 - 54
Reference Cited: 31
Number of Pages: vii+89
Inspec Accession Number: 6319884

Abstract:
We present an **illumination**-based method for synthesizing **images** of an ob novel **viewing** conditions. Our method requires as few as three **images** of th taken under variable **illumination**, but from a fixed viewpoint. Unlike multi-v **image** synthesis, our method does not require the determination of point or l correspondences. Furthermore, our method is able to synthesize not simply n viewpoints, but novel **illumination** conditions as well. We demonstrate the e of our approach by generating synthetic **images** of human **faces**

Index Terms:
computational geometry computer graphics image processing lighting fixed viewp faces illumination-based image synthesis lighting pose synthetic images view conditions

Documents that cite this document
There are no citing documents available in IEEE Xplore at this time.

Search Results [PDF FULL-TEXT 1288 KB] PREV NEXT DOWNLOAD CITATION

Home | Log-out | Journals | Conference Proceedings | Standards | Search by Author | Basic Search | Advanced Search | Join IEEE | Web Account | New this week | OPAC Linking Information | Your Feedback | Technical Support | Email Alerting | No Robots Please | Release Notes | IEEE Online Publications | Help | FAQ| Terms | Back to Top

Illumination-Based Image Synthesis: Creating Novel Images of Human Faces Under Differing Pose and Lighting*

Athinodoros S. Georghiades   Peter N. Belhumeur   David J. Kriegman

Center for Computational Vision and Control
Yale University
New Haven, CT 06520-8267

Beckman Institute
University of Illinois, Urbana-Champaign
Urbana, IL 61801

Abstract

We present an illumination-based method for synthesizing images of an object under novel viewing conditions. Our method requires as few as three images of the object taken under variable illumination, but from a fixed viewpoint. Unlike multi-view based image synthesis, our method does not require the determination of point or line correspondences. Furthermore, our method is able to synthesize not simply novel viewpoints, but novel illumination conditions as well. We demonstrate the effectiveness of our approach by generating synthetic images of human faces.

1 Introduction

We present an illumination-based method for creating novel images of an object under differing pose and lighting. This method uses as few as three images of the object taken under variable lighting but fixed pose to estimate the object's albedo and generate its geometric structure. Our approach does not require any knowledge about the light source directions in the modeling images, or the establishment of point or line correspondences.

In contrast, nearly all approaches to view synthesis or image-based rendering take a set of images gathered from multiple viewpoints and apply techniques akin to structure from motion [17, 28, 6], stereopsis [21, 9], image transfer [3], image warping [18, 20, 24], or image morphing [7, 23]. Each of these methods requires the establishment of correspondence between image data (e.g. pixels) across the set. (Unlike other methods, the Lumigraph [12, 19] exhaustively samples the ray space and renders images of an object from novel viewpoints by taking $2-D$ slices of the $4-D$ light field at the appropriate directions.) Since dense correspondence is difficult to obtain, most methods extract sparse image features (e.g. corners, lines), and may use multi-view geometric constraints (e.g. the trifocal tensor [2, 1]) or scene-dependent geometric constraints [9, 8] to reduce the search process and constrain the estimates. By using a sequence of images taken at nearby viewpoints, incremental tracking can further simplify the process, particularly when features are sparse.

For these approaches to be effective, there must be sufficient texture or viewpoint-independent scene features, such as albedo discontinuities or surface normal discontinuities. From sparse correspondence, the epipolar geometry can be established and stereo techniques can be used to provide dense reconstruction. Underlying nearly all such stereo algorithms is a constant brightness assumption that is, the intensity (irradiance) of corresponding pixels should be the same. In turn, constant brightness implies two seldom stated assumptions: (1) The scene is Lambertian, and (2) the lighting is static with respect to the scene only the viewpoint is changing.

In the presented illumination-based approach, we also assume that the surface is Lambertian, although this assumption is very explicit. As a dual to the second point listed above, our method requires that the camera remains static with respect to the scene - only the lighting is changing. As a consequence, geometric correspondence is trivially established, and so the method can be applied to scenes where it is difficult to establish multi-viewpoint correspondence, namely scenes that are highly textured (i.e. where image features are not sparse) or scenes that completely lack texture (i.e. where there are insufficient image features).

At the core of our approach for generating novel viewpoints is a variant of photometric stereo [27, 29, 14, 13, 30] which simultaneously estimates geometry and albedo across the scene. However, the main limitation of classical photometric stereo is that the light source positions must be accurately known, and this necessitates a fixed lighting rig as might be possible in an industrial setting. Instead, the proposed method *does not* require knowledge of light source locations, and so illumination could be varied by simply waiving a light around the scene.

In fact, our method derives from work by Belhumeur and Kriegman in [5] where they showed that a small set of images with unknown light source directions can

*P. N. Belhumeur and A. S. Georghiades were supported by a Presidential Early Career Award, an NSF Career Award IRI-9703134, and ARO grant DAAH04-95-1-0494. D. J. Kriegman was supported by NSF under NYI, IRI-9257990, and by ARO grant DAAG55-98-1-0168.

be used to generate a representation the illumination cone – which models the complete set of images of an object (in fixed pose) under all possible illumination. This method had as its pre-cursor the work of Shashua [25] who showed that, in the absence of shadows, the set of images of an object lies in a $3 - D$ subspace in the image space. Generated images from the illumination cone representation accurately depict shading and attached shadows under extreme lighting; in [11] the cone representation was extended to include cast shadows (shadows the object casts on itself) for objects with non-convex shapes. Unlike attached shadows, cast shadows are global effects, and their prediction requires the reconstruction of the object's surface.

In generating the geometric structure, multi-viewpoint methods typically estimate depth directly from corresponding image points [21, 9]. It is well known that without sub-pixel correspondence, stereopsis provides a modest number of disparities over the effective operating range, and so smoothness or regularization constraints are used to interpolate and provide smooth surfaces. The presented illumination-based method estimates surface normals which are then integrated to generate a surface. As a result, very subtle changes in depth are recovered as demonstrated in the synthetic images in Figures 4 and 5. Those images show also the effectiveness of our approach in generating realistic images of faces under novel pose and illumination conditions.

2 Illumination Modeling

In [5], Belhumeur and Kriegman have shown that, for a convex object with a Lambertian reflectance function, the set of all images under an arbitrary combination of point light sources forms a convex polyhedral cone in the image space $\mathbb{R}^n$ which can be constructed with as few as three images.

Let $\mathbf{x} \in \mathbb{R}^n$ denote an image with n pixels of a convex object with a Lambertian reflectance function illuminated by a single point source at infinity. Let $B \in \mathbb{R}^{n \times 3}$ be a matrix where each row in B is the product of the albedo with the inward pointing unit normal for a point on the surface projecting to a particular pixel in the image. A point light source at infinity can be represented by $\mathbf{s} \in \mathbb{R}^3$ signifying the product of the light source intensity with a unit vector in the direction of the light source. A convex Lambertian surface with normals and albedo given by B, illuminated by $\mathbf{s}$, produces an image $\mathbf{x}$ given by

$$\mathbf{x} = \max(B\mathbf{s}, \mathbf{0}), \qquad (1)$$

where $\max(B\mathbf{s}, \mathbf{0})$ sets to zero all negative components of the vector $B\mathbf{s}$. The pixels set to zero correspond to the surface points lying in an attached shadow. Convexity of the object's shape is assumed at this point to avoid cast shadows. It should be noted that when no part of the surface is shadowed, $\mathbf{x}$ lies in the 3-D subspace $\mathcal{L}$ given by the span of the columns of B.

If an object is illuminated by k light sources at infinity, then the image is given by the superposition of the images which would have been produced by the individual light sources, i.e.,

$$\mathbf{x} = \sum_{i=1}^{k} \max(B\mathbf{s}_i, \mathbf{0}) \qquad (2)$$

where $\mathbf{s}_i$ is a single light source. Due to the inherent superposition, it follows that the set of all possible images $\mathcal{C}$ of a convex Lambertian surface created by varying the direction and strength of an arbitrary number of point light sources at infinity is a convex cone. It is also evident from Equation 2 that this convex cone is completely described by matrix B.

This suggests a way to construct the illumination model for an individual: gather three or more images of the face without shadowing illuminated by a single light source at unknown locations but viewed under fixed pose, and use them to estimate the three-dimensional illumination subspace $\mathcal{L}$. This can be done by first normalizing the images to unit length and then estimating the best three-dimensional orthogonal basis B^* using a least-squares minimization technique such as singular value decomposition (SVD). Note that the basis B^* differs from B by an unknown linear transformation, i.e., $B = B^*A$ where $A \in GL(3)$ [10, 13, 22]; for any light source $\mathbf{s}$, $\mathbf{x} = B\mathbf{s} = (B^*A)(A^{-1}\mathbf{s})$. Nevertheless, both B^* and B define the same illumination cone and represent valid illumination models.

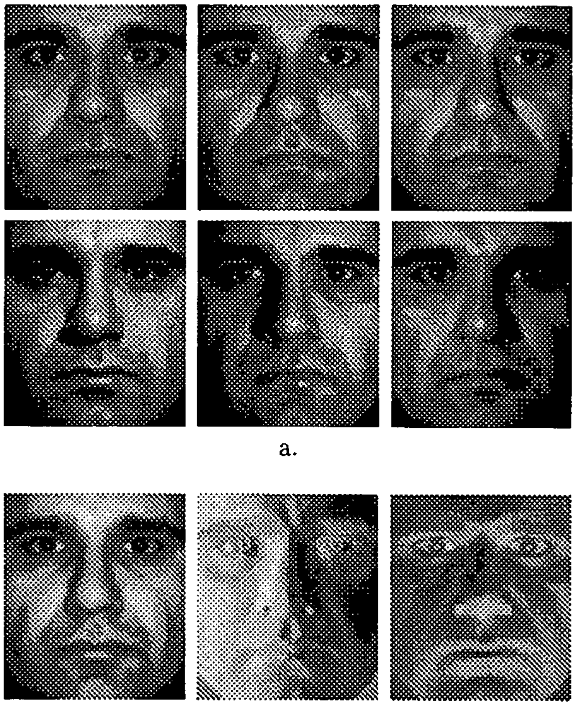
Unfortunately, using SVD in the above procedure leads to an inaccurate estimate of B^*. For even a convex object whose Gaussian image covers the Gauss sphere, there is only one light source direction (the viewing direction) for which no point on the surface is in shadow. For any other light source direction, shadows will be present. If the object is non-convex, such as a face, then shadowing in the modeling images is likely to be more pronounced. When SVD is used to find B^* from images with shadows, these systematic errors bias its estimate significantly. Therefore, an alternative way is needed to find B^* that takes into account the fact that some data values should not be used in the estimation.

We have implemented a variation of [26] (see also [28, 16]) that finds a basis B^* for the 3-D linear subspace $\mathcal{L}$ from image data with missing elements. To begin, define the data matrix for c images of an individual to be $X = [\mathbf{x}_1 \ldots \mathbf{x}_c]$. If there were no shadowing, X would be rank 3 (assuming no image noise), and we could use SVD to factorize X into $X = B^*S^*$ where S^* is a $3 \times c$ matrix the columns of which are the light source directions scaled by the light intensities $\mathbf{s}_i$ for all c images.

Since the images have shadows (both cast and attached), and possibly saturations, we first have to determine which data values are invalid. Unlike saturations which can be trivially determined, finding shadows is more involved. In our implementation, a pixel is assigned to be in shadow if its value divided by its corresponding albedo is below a threshold. As an initial estimate of the albedo, we use the average of the modeling (or training) images. A conservative threshold is then chosen to determine shadows making it almost certain no invalid data is included in the estimation process, at the small expense of throwing away some valid data. After finding the invalid data, the following estimation method is used: without doing any row or column permutations sift out all the full rows (with no invalid data) of matrix X to form a full sub-matrix $\tilde{X}$. Note that the number of pixels in an image (i.e. the number of rows of X) is much larger than the number of images (i.e. the number of columns of X), which means we can always find a large number of full rows so that the number of rows of $\tilde{X}$ is larger than its number of columns. Therefore, perform SVD on $\tilde{X}$ to get a fairly good initial estimate of S^*. Fix S^* and estimate each of the rows of B^* independently using least squares. Then, fix B^* and update each of the light source direction $\mathbf{s}_i$ independently, again using least squares. Repeat these last two steps until estimates converge. In our experiments, the algorithm is very well behaved, converging to the global minimum within 10-15 iterations. Though it is possible to converge to a local minimum, we never observed this either in simulation or in practice.

Figure 1 demonstrates the process for constructing the illumination model. Figure 1.a shows six of the original single light source images of a face used in the estimation of B^*. Note that the light source in each image moves only by a small amount ($\pm 15^o$ in either direction) about the viewing axis. Despite this, the images do exhibit shadowing, e.g. left and right of the nose. In fact, there is a tradeoff in the image acquisition process: the smaller the motion of the light source, meaning fewer shadows present in the images, the worse the conditioning of the estimation problem. If, on the other hand, the light source moves excessively, despite the improvement in the conditioning, more extensive shadowing can increase the possibility of having too few (less than three) valid measurements with a fixed number of images for some parts of the face. Therefore, the light source should move in moderation as in the images shown in Figure 1.a.

Figure 1.b shows the basis images of the estimated matrix B^*. These basis images encode not only the albedo (reflectance) of the face but also its surface normal field. They can be used to construct images of the face under arbitrary and quite extreme illumination conditions. However, the image formation model in Equation 1 does not account for cast shadows of non-convex objects such as faces. In order to determine which parts of the image are in cast shadows, given a light source direction, we need to reconstruct the surface of the face (see next section) and then use ray-tracing techniques.


a.

b.

Figure 1: a) Six of the original single light source images used to estimate B^*. Note that the light source in each image moves only by a small amount ($\pm 15^o$ in either direction) about the viewing axis. Despite this, the images do exhibit shadowing. b) The basis images of B^*.

3 Surface Reconstruction

In this section, we demonstrate how we can generate an object's surface from B^* after enforcing the integrability constraint on the surface normal field. It has been shown [4, 31] that from multiple images, in which the light source directions are unknown, one can only recover a Lambertian surface up to a three-parameter family given by the generalized bas-relief (GBR) transformation. This family scales the relief (flattens or extrudes) and introduces an additive plane. It has also been shown that the family of GBR transformations is the only one that preserves integrability.

3.1 Enforcing Integrability

The vector field B^* estimated in Section 2 may not be integrable, i.e., it may not correspond to a smooth surface. So, prior to reconstructing the surface up to GBR, the integrability constraint must be enforced on B^*. Since no method has been developed to enforce the integrability during the estimation of B^*, we enforce it afterwards. That is, given B^* estimate a matrix $A \in GL(3)$ such that $\hat{B} = B^*A$ corresponds to an integrable normal field; the development follows [31].

Consider a continuous surface defined as the graph of $z(x, y)$, and let $\mathbf{b}(x, y)$ be the corresponding normal field scaled by an albedo field. The integrability constraint for a surface is $z_{xy} = z_{yx}$ where subscripts denote partial derivatives. In turn, $\mathbf{b}(x, y)$ must satisfy:

$$\left(\frac{b_1}{b_3}\right)_y = \left(\frac{b_2}{b_3}\right)_x$$

To estimate A such that $\mathbf{b}^T(x, y) = \mathbf{b}^{*^T}(x, y)A$, we expand this out. Letting the columns of A be denoted by A_1, A_2, A_3 yields

$$(\mathbf{b}^{*^T} A_3)(\mathbf{b}_x^{*^T} A_2) - (\mathbf{b}^{*^T} A_2)(\mathbf{b}_x^{*^T} A_3) = \\ (\mathbf{b}^{*^T} A_3)(\mathbf{b}_y^{*^T} A_1) - (\mathbf{b}^{*^T} A_1)(\mathbf{b}_y^{*^T} A_3)$$

which can be expressed as

$$\mathbf{b}^{*T} S_1 \mathbf{b}_x^* = \mathbf{b}^{*T} S_2 \mathbf{b}_y^* \tag{3}$$

where $S_1 = A_3 A_2^T - A_2 A_3^T$ and $S_2 = A_3 A_1^T - A_1 A_3^T$.

S_1 and S_2 are skew-symmetric matrices and have three degrees of freedom. Equation 3 is linear in the six elements of S_1 and S_2. From the estimate of B^* discrete approximations of the partial derivatives ($\mathbf{b}_x^*$ and $\mathbf{b}_y^*$) are computed, and then SVD is used to solve for the six elements of S_1 and S_2. In [31], it was shown that the elements of S_1 and S_2 are cofactors of A, and a simple method for computing A from the cofactors was presented. This procedure only determines six degrees of freedom of A. The other three correspond to the GBR transformation [4] and can be chosen arbitrarily because a GBR transformation preserves integrability. The surface corresponding to $\hat{B} = B^*A$ differs from the true surface by GBR, i.e., $\hat{z}(x, y) = \lambda z(x, y) + \mu x + \nu y$ for arbitrary λ, μ, ν with $\lambda \neq 0$.

3.2 Generating a GBR surface

After enforcing integrability, we can now reconstruct the corresponding surface $\hat{z}(x, y)$. Note that $\hat{z}(x, y)$ is not a Euclidean reconstruction of the face, but a representative element of the orbit under a GBR transformation. Despite this, both the shading *and* the shadowing will be correct for images synthesized from such a surface [4].

To find $\hat{z}(x, y)$, we use the variational approach presented in [15]. A surface $\hat{z}(x, y)$ is fit to the given components of the gradient p and q by minimizing the functional

$$\int\int_\Omega (\hat{z}_x - p)^2 + (\hat{z}_y - q)^2 \, dx \, dy.$$

the Euler equation of which reduces to $\nabla^2 z = p_x + q_y$. By enforcing the right natural boundary conditions and employing an iterative scheme that uses a discrete approximation of the Laplacian, we can reconstruct the surface $\hat{z}(x, y)$ [15].

Recall that a GBR transformation scales the relief (flattens or extrudes) and introduces an additive plane. To resolve this GBR ambiguity, we take advantage of the fact that we are dealing with human faces which constitute a well known class of objects. We can therefore exploit the left-to-right symmetry of faces and the fairly constant ratios of distances between facial features such as the eyes, the nose, and the forehead. (In the case when the class of objects is not well defined, the issue of resolving the GBR ambiguity becomes more subtle and is essentially an open problem.) A surface of a face that has undergone a GBR transformation will have different distance ratios and can be asymmetric. These differences allow us to estimate the three parameters of the GBR transformation which we can then invert. Note that this inverse transformation is applied to both the estimated surface $\hat{z}(x, y)$ and $\hat{B}$. Even though this inverse operation (which is also a GBR transformation) may not completely resolve the ambiguity of the relief because of errors in the estimation of the GBR parameters, it nevertheless comes very close to that effect. After all, our purpose is not to reconstruct the exact Euclidean surface of the face, but to create realistic images of a face under differing pose and illumination. Moreover, since shadows are preserved under GBR transformations [4], images synthesized under an arbitrary light source from a surface whose normal field has been GBR transformed will have correct shadowing. This means that the residual GBR transformation (after resolving the ambiguity) will not affect the image synthesis with variable illumination.

Figure 2 shows the reconstructed surface of the face shown in Figure 1 after resolving the GBR ambiguity. The first basis image of B^* shown in Figure 1.b has been texture-mapped on the surface. Even though we cannot recover the exact Euclidean structure of the face (i.e. resolve the ambiguity completely), we can still generate synthetic images of a face under variable pose where the shape distortions due to the residual GBR ambiguity are quite small and not visually detectable.

4 Image Synthesis

We first demonstrate the ability of our method to generate images of an object under novel illumination conditions but fixed pose. Figure 3 shows sample single light source images of a face generated with the image formation model in Equation 1 which has been extended to account for cast shadows. To determine cast shadows, we employ ray-tracing that uses the reconstructed surface of the face $\hat{z}(x, y)$ after resolving the GBR ambiguity. Specifically, a point on the surface is in cast shadow if, for a given light source direction, a ray emanating from that point parallel to the light source direction intersects the surface at some other point. With this extended image formation model, the generated images exhibit realistic shading and, despite the small presence of shadows in the images in Figure 1.a, have strong attached and cast shadows.

Figure 4 displays a set of synthesized images of the

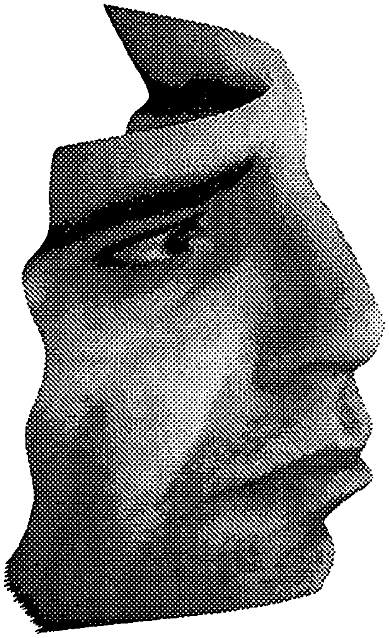


Figure 2: The reconstructed surface.

the face viewed under variable pose but with fixed lighting. The images were created by rigidly rotating the reconstructed surface shown in Figure 2 first about the horizontal and then about the vertical axis. Along the rows from left to right, the azimuth varies (in 10 degree intervals) from 30 degrees to the right of the face to 10 degrees to the left. Down the columns, the elevation varies (again in 10 degree intervals) from 20 degrees above the horizon to 30 degrees below. For example, in the bottom image of the second column from the left the surface has an azimuth of 20 degrees to the right and an elevation of 30 degrees below the horizon. The single light source is following the face around as it changes pose. This implies that a patch on the surface has the same intensity in all poses. It is interesting to see that the images look quite realistic with maybe the exception of the three right images in the bottom row which appear to be a little flattened. This is not due to any errors during the geometric or photometric modeling but probably due to our visual priors; we are not used to looking at a face from above.

In Figure 5, we combine both variations in viewing conditions to synthesize images of the face under novel pose and lighting. We used the same poses as in Figure 4 but now the light from the single point source is fixed to come along the gaze direction of the face in the top-right image. Therefore, as the face moves around and its gaze direction changes with respect to the light source direction, the shading of the surface changes and both attached and cast shadows are formed, as one would expect. The synthesized images seem to agree with our visual intuition.

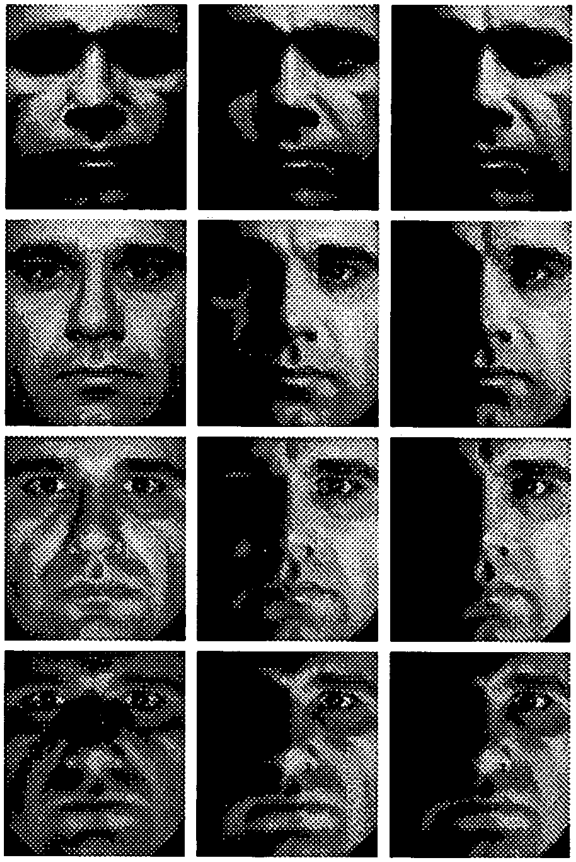


Figure 3: Sample images of the face under novel illumination conditions but fixed pose.

5 Discussion

Appearance variation of an object caused by small changes in illumination under fixed pose can provide enough information to estimate (under the assumption of a Lambertian reflectance function) the object's surface normal field scaled by its albedo. In the presented method, as few as three images with no knowledge of the light source directions can be used in the estimation. The estimated surface normal field can then be integrated to reconstruct the object's surface. Unlike multi-view based image synthesis, our approach does not require the determination of point or line correspondences to do the surface reconstruction. Since we are dealing with a well known class of objects, we can acceptably resolve the GBR ambiguity of the reconstructed surface. Then, the surface together with the surface normal field scaled by the albedo are sufficient for synthesizing images of the object under novel pose and lighting.

The effectiveness of our approach stems from three reasons. First, the estimation of the illumination model B^* does not use any invalid data (such as shadows) which would otherwise lead to large biases. Second, the integrability constraint is enforced on the surface normal field which significantly improves the surface reconstruction. Last, unlike classical photometric stereo, our method requires no knowledge of light source locations. This obviates the need of error-prone calibration of a fixed lighting rig where any errors in estimating the position of the light sources can propagate to the estimation of the illumination model causing large inaccuracies. These reasons have to led to improved performance and we have demonstrated this by synthesizing realistic images of human faces.

References

[1] S. Avidan, T. Evgeniou, A. Shashua, and T. Poggio. Image-based view synthesis by combining trilinear tensors and learning techniques. In *ACM Symposium on Virtual Reality Software and Technology*, 1997.

[2] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 1034–1040, 1997.

[3] E. Barett, M. Brill, N. Haag, and P. Payton. Invariant linear methods in photogrammetry and model matching. In J. Mundy and A. Zisserman, editors, *Geometric Invariance in Computer Vision*, pages 277 292. MIT Press, 1992.

[4] P. Belhumeur, D. Kriegman, and A. Yuille. The bas-relief ambiguity. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 1040–1046, 1997.

[5] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 270–277, 1996.

[6] R. Carceroni and K. Kutulakos. Shape and motion of 3-d curves from multi-view image scenes. In *Image Understanding Workshop*, pages 171 176, 1998.

[7] S. Chen and L. Williams. View interpolation for image synthesis. In *Computer Graphics (SIGGRAPH)*, pages 279 288, 1993.

[8] G. Chou and S. Teller. Multi-image correspondence using geometric and structural constraints. In *Image Understanding Workshop*, pages 869 874, 1997.

[9] P. Debevec, C. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Computer Graphics (SIGGRAPH)*, pages 11–20, 1996.

[10] R. Epstein, A. Yuille, and P. N. Belhumeur. Learning and recognizing objects using illumination subspaces. In *Proc. of the Int. Workshop on Object Representation for Computer Vision*, 1996.

[11] A. Georghiades, D. Kriegman, and P. Belhumeur. Illumination cones for recognition under variable lighting: Faces. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 1998.

[12] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The Lumigraph. In *Computer Graphics (SIGGRAPH)*, pages 43–54, 1996.

[13] H. Hayakawa. Photometric stereo under a light-source with arbitrary motion. *JOSA-A*, 11(11):3079 3089, Nov. 1994.

[14] B. Horn. *Computer Vision*. MIT Press, Cambridge, Mass., 1986.

[15] B. Horn and M. Brooks. The variational approach to shape from shading. *Computer Vision, Graphics and Image Processing*, 35:174–208, 1992.

[16] D. Jacobs. Linear fitting with missing data: Applications to structure from motion and characterizing intensity images. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, 1997.

[17] J. Koenderink and A. Van Doorn. Affine structure from motion. *JOSA-A*, 8(2):377 385, 1991.

[18] S. Laveau and O. Faugeras. 3-D scene representation as a collection of images and fundamental matrices. Technical Report 2205, INRIA-Sophia Antipolis, February 1994.

[19] M. Levoy and P. Hanrahan. Light field rendering. In *Computer Graphics (SIGGRAPH)*, pages 31 42, 1996.

[20] W. R. Mark, L. McMillan, and G. Bishop. Post-rendering 3d warping. In *Computer Graphics (SIGGRAPH)*, pages 39–46, 1997.

[21] L. Matthies, R. Szeliski, and T. Kanade. Kalman filter-based algorithms for estimating depth from image sequences. *Int. J. Computer Vision*, 3:293–312, 1989.

[22] R. Rosenholtz and J. Koenderink. Affine structure and photometry. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 790 795, 1996.

[23] S. Seitz and C. Dyer. View morphing. In *Computer Graphics (SIGGRAPH)*, pages 21–30, 1996.

[24] J. Shade, S. Gortler, L. wei He, and R. Szeliski. Layered depth maps. In *Computer Graphics (SIGGRAPH)*, pages 251–258, 1998.

[25] A. Shashua. *Geometry and Photometry in 3D Visual Recognition*. PhD thesis, MIT, 1992.

[26] H. Shum, K. Ikeuchi, and R. Reddy. Principal component analysis with missing data and its application to polyhedral object modeling. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 17(9):854 867, September 1995.

[27] W. Silver. *Determining Shape and Reflectance Using Multiple Images*. PhD thesis, MIT, Cambridge, MA, 1980.

[28] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Computer Vision*, 9(2):137–154, 1992.

[29] R. Woodham. Analysing images of curved surfaces. *Artificial Intelligence*, 17:117–140, 1981.

[30] Y. Yu and J. Malik. Recovering photometric properties of architectural scenes from photographs. In *Computer Graphics (SIGGRAPH)*, pages 207–218, 1998.

[31] A. Yuille and D. Snow. Shape and albedo from multiple images using integrability. In *Proc. IEEE Conf. on Comp. Vision and Patt. Recog.*, pages 158 164, 1997.

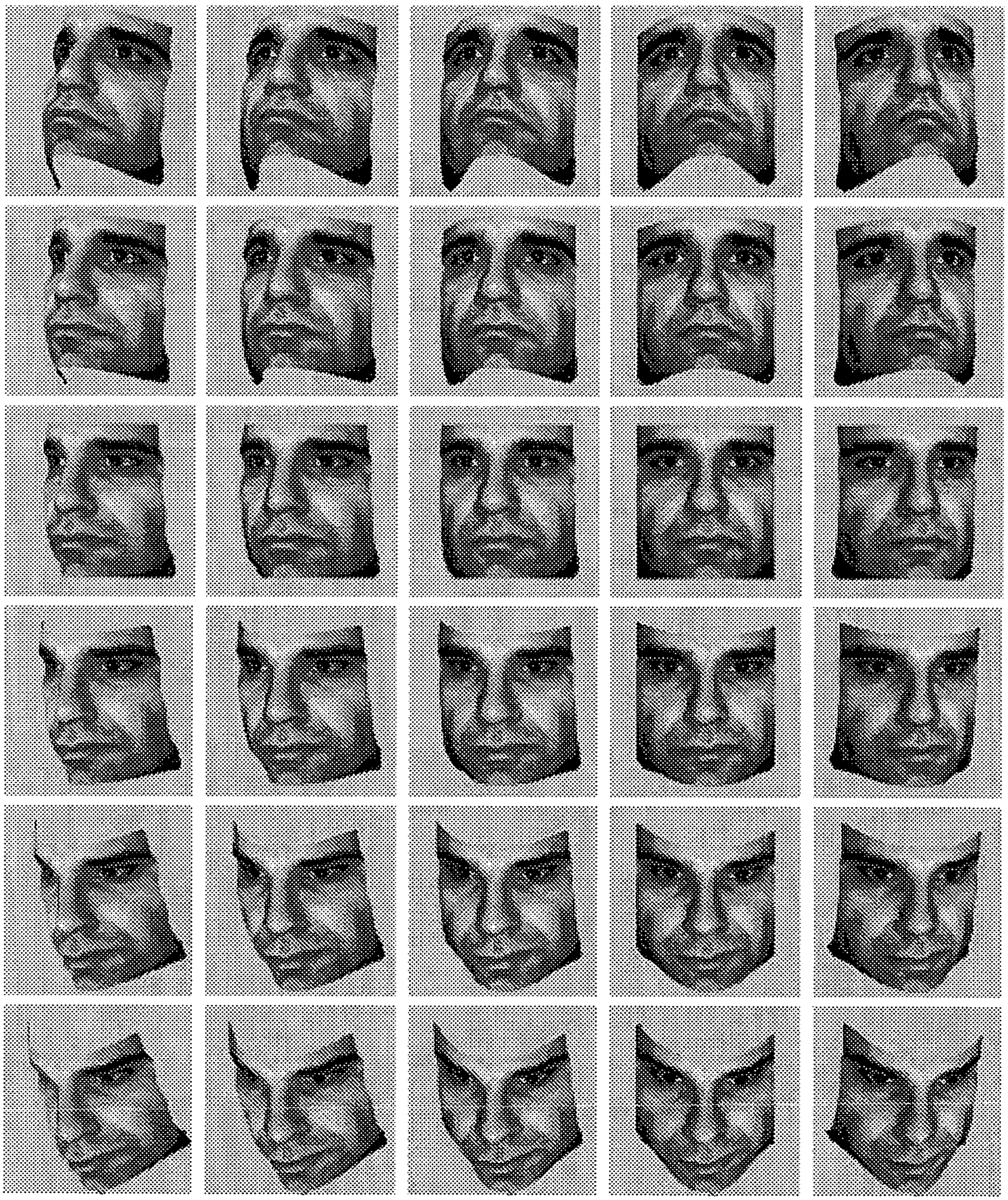


Figure 4: Synthesized images under variable pose but with fixed lighting; the single light source is following the face.

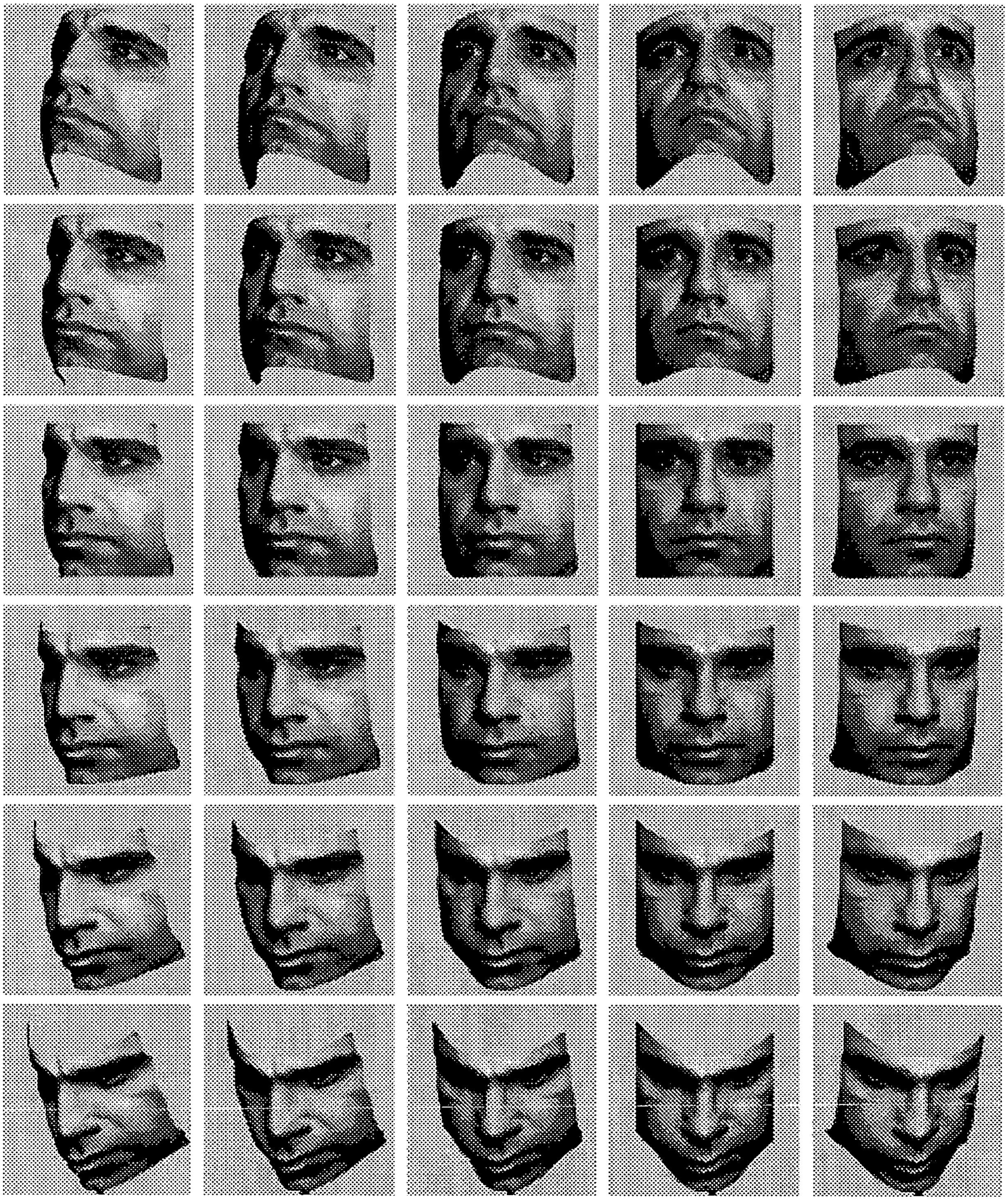


Figure 5: Synthesized images under *both* variable pose and lighting. As the face moves around the single light source stays fixed resulting to image variability due to changes in pose and illumination conditions.

IEEE HOME | SEARCH IEEE | SHOP | WEB ACCOUNT | CONTACT IEEE

IEEE

Membership Publications/Services Standards Conferences Careers/Jobs

IEEE Xplore® RELEASE 1.6

Welcome
United States Patent and Trademark Office

Help FAQ Terms IEEE Peer Review Quick Links » ABS

Welcome to IEEE Xplore®
- Home
- What Can I Access?
- Log-out

Tables of Contents
- Journals & Magazines
- Conference Proceedings
- Standards

Search
- By Author
- Basic
- Advanced

Member Services
- Join IEEE
- Establish IEEE Web Account
- Access the IEEE Member Digital Library

Search Results [PDF FULL-TEXT 280 KB] PREV NEXT DOWNLOAD CITATION

Request Permissions RIGHTS LINK

View-based active appearance models

Cootes, T.F. Walker, K. Taylor, C.J.
Dept. of Imaging Sci. & Biomed. Eng., Manchester Univ., UK;

This paper appears in: **Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on**

Meeting Date: 03/28/2000 - 03/30/2000
Publication Date: 28-30 March 2000
Location: Grenoble France
On page(s): 227 - 232
Reference Cited: 15
Number of Pages: xiv+560
Inspec Accession Number: 6577281

Abstract:

We demonstrate that a small number of 2D statistical models are sufficient to shape and appearance of a face from any viewpoint (full profile to front-to-pa model is linear and can be matched rapidly to new images using the active ap model algorithm. We show how such a set of models can be used to estimate to track faces through large angles of head rotation and to synthesize faces fr viewpoints

Index Terms:

face recognition image matching statistical analysis tracking 2D statistical models capturing face synthesis face tracking head pose estimation head rotation image linear model shape capturing unseen viewpoints view-based active appearance mod

Documents that cite this document

Select link to view other documents in the database that cite this one.

Search Results [PDF FULL-TEXT 280 KB] PREV NEXT DOWNLOAD CITATION

Home | Log-out | Journals | Conference Proceedings | Standards | Search by Author | Basic Search | Advanced Search | Join IEEE | Web Account | New this week | OPAC Linking Information | Your Feedback | Technical Support | Email Alerting | No Robots Please | Release Notes | IEEE Online Publications | Help | FAQ| Terms | Back to Top

Copyright © 2004 IEEE — All rights reserved

eee e eee ea c c ab ac ?a be

View-Based Active Appearance Models

T.F. Cootes, K. Walker, C.J. Taylor
Dept. Imaging Science and Biomedical Engineering
University of Manchester, Manchester M13 9PT U.K.
t.cootes@man.ac.uk

Abstract

We demonstrate that a small number of 2D statistical models are sufficient to capture the shape and appearance of a face from any viewpoint (full profile to fronto-parallel). Each model is linear and can be matched rapidly to new images using the Active Appearance Model algorithm. We show how such a set of models can be used to estimate head pose, to track faces through large angles of head rotation and to synthesize faces from unseen viewpoints.

1 Introduction

The appearance of a face in a 2D image can change dramatically as the viewing angle changes. The majority of work on face tracking and recognition assumes near fronto-parallel views, and tends to break down when presented with large rotations or profile views. Three general approaches have been used to deal with this; a) use a full 3D model [15], b) introduce nonlinearities into a 2D model [6] and c) use a set of models to represent appearance from different view points [11]. In this paper we explore the last approach, using statistical models of shape and appearance to represent the variations in appearance from a particular viewpoint.

These appearance models are trained on example images labelled with sets of landmarks to define the correspondences between images [1]. Lanitis *et. al.*[9] showed that a linear model was sufficient to simulate considerable changes in viewpoint, as long as all the modelled features (the landmarks) remained visible. A model trained on near fronto-parallel face images can cope with pose variations of up to 45^o either side. For much larger angle displacements, some features become occluded, and the assumptions of the model break down.

We demonstrate that to deal with full 180^o rotation (from left profile to right profile), we need only 5 models, roughly centred on viewpoints at -90^o,-45^o,0^o,45^o,90^o (where 0^o corresponds to fronto-parallel). The pairs of models at $\pm 90^o$ (full profile) and $\pm 45^o$(half profile) are simply reflections of each other, so there are only 3 distinct models. We can use these models for estimating head pose, for tracking faces through wide changes in orientation and for synthesizing new views of a subject given a single view.

Each model is trained on labelled images of a variety of people with a range of orientations chosen so none of the features for that model become occluded. The different models use different sets of features (see Figure 1). Each example view can then be approximated using the appropriate appearance model with a vector of parameters, **c**. We assume that as the orientation changes, the parameters, **c**, trace out an approximately elliptical path. We can learn the relationship between **c** and head orientation, allowing us to both estimate the orientation of any head and to be able to synthesize a face at any orientation.

By using the Active Appearance Model algorithm [4, 1] we can match any of the individual models to a new image rapidly. If we know in advance the approximate pose, we can easily select the most suitable model. If we do not know, we can search with each of the five models and choose the one which achieves the best match. Once a model is selected and matched, we can estimate the head pose, and thus track the face, switching to a new model if the head pose varies significantly.

Given a single image of a new person, we can match the models to estimate the pose. We can then use the best fitting model to generate new views from angles similar to that of the original image. We can also exploit correlations across models of different views to estimate the appearance of the subject in a completely different view. Though this can perhaps be done most effectively with a full 3D model [15], we demonstrate that good results can be achieved just with a set of 2D models.

In the following we describe the techniques in more detail and give examples of the model, its ability to estimate pose, to track faces and to synthesize unseen views.

2 Background

Statistical models of shape and texture have been widely used for recognition, tracking and synthesis [7, 9, 4, 14], but have tended to only be used with near fronto-parallel images.

Moghaddam and Pentland [11] describe using view-based eigenface models to represent a wide variety of viewpoints. Our work is similar to this, but by including shape variation (rather than the rigid eigen-patches), we require fewer models and can obtain better reconstructions with fewer model modes.

Maurer and von der Malsburg [10] demonstrated tracking heads through wide angles by tracking graphs whose nodes are facial features, located with Gabor jets. The system is effective for tracking, but is not able to synthesize the appearance of the face being tracked.

Murase and Nayar [6] showed that the projections of multiple views of a rigid object into an eigenspace fell on a 2D manifold in that space. By modelling this manifold they could recognise objects from arbitrary views. A similar approach has been taken by Gong *et. al.*[13, 8] who use non-linear representations of the projections into an eigen-face space for tracking and pose estimation, and by Graham and Allinson [5] who use it for recognition from unfamiliar viewpoints.

Romdhani *et. al.*[12] has extended the Active Shape Model to deal with full 180° rotation of a face using a non-linear model. However, the non-linearities mean the method is slow to match to a new image.

Vetter [15] has demonstrated how a 3D statistical model of face shape and texture can be used to generate new views given a single view. The model can be matched to a new image from more or less any viewpoint using a general optimisation scheme, though this is slow. By explicitly taking into account the 3D nature of the problem, this approach is likely to yield better reconstructions than the purely 2D method described below. However, the view based models we propose could be used to drive the parameters of the 3D head model, speeding up matching times.

3 Statistical Models of Appearance

An appearance model can represent both the shape and texture variability seen in a training set. The training set consists of labelled images, where key landmark points are marked on each example object. The training set is usually labelled manually, though automatic methods are being developed. For instance, Figure 1 shows examples of labelled images used to train the view-based face models.

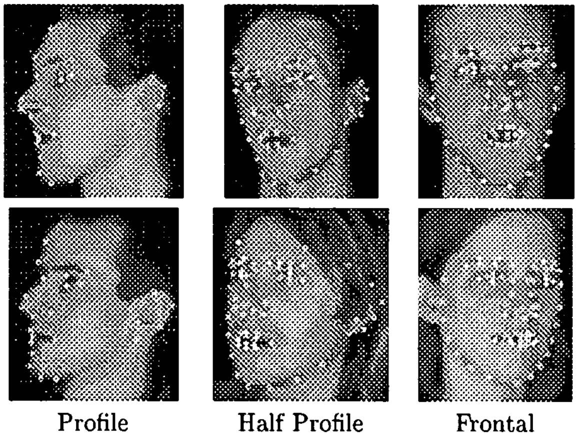

Profile Half Profile Frontal

Figure 1. Examples from the training sets for the models

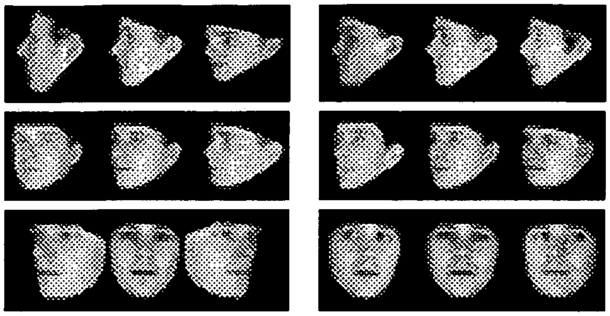
Given such a set we can generate a statistical models of shape and texture variation (see [1, 4] for details). The shape of an object can be represented as a vector $\mathbf{x}$ and the texture (grey-levels or colour values) represented as a vector $\mathbf{g}$. The appearance model has parameters, $\mathbf{c}$, controlling the shape and texture according to

$$\begin{aligned} \mathbf{x} &= \bar{\mathbf{x}} + \mathbf{Q}_s\mathbf{c} \\ \mathbf{g} &= \bar{\mathbf{g}} + \mathbf{Q}_g\mathbf{c} \end{aligned} \quad (1)$$

where $\bar{\mathbf{x}}$ is the mean shape, $\bar{\mathbf{g}}$ the mean texture and $\mathbf{Q}_s$,$\mathbf{Q}_s$ are matrices describing the modes of variation derived from the training set.

We trained three distinct models on data similar to that shown in Figure 1. The profile model was trained on 234 landmarked images taken of 15 individuals from different orientations. The half-profile model was trained on 82 images, and the frontal model on 294 images.

An example image can be synthesised for a given $\mathbf{c}$ by generating a texture image from the vector $\mathbf{g}$ and warping it using the control points described by $\mathbf{x}$. For instance, Figure 2 shows the effects of varying the first two appearance model parameters, c_1, c_2, of models trained on a set of face images, labelled as shown in Figure 1. These change both the shape and the texture component of the synthesised image.



c_1 varies ±2 s.d.s c_2 varies ±2 s.d.s

Figure 2. First two modes of the face models (top to bottom: profile, half-profile and frontal)

4 Predicting Pose

We assume that the model parameters are related to the viewing angle, θ, approximately as

$$\mathbf{c} = \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \qquad (2)$$

where $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ are vectors estimated from training data (see below).

(Here we consider only rotation about a vertical axis - head turning. Nodding can be dealt with in a similar way.)

This is an accurate representation of the relationship between the shape, $\mathbf{x}$, and orientation angle under an affine projection (the landmarks trace circles in 3D which are projected to ellipses in 2D), but our experiments suggest it is also an acceptable approximation for the appearance model parameters, $\mathbf{c}$.

In order to learn the relationship for a given model, we must know the orientation of each of our training examples. We do not yet have access to a system which can measure it accurately, such as that used by [12, 8, 13]. However, we are able to estimate the angle by finding the frames in our training sequences at full profile and fronto-parallel by eye, then assuming a constant rate of rotation across the frames between. This leads to images labelled with orientations, θ_i, accurate to about ±10°. For each such image we find the best fitting model parameters, $\mathbf{c}_i$. We then perform regression between $\{\mathbf{c}_i\}$ and the vectors $\{(1, \cos(\theta_i), \sin(\theta_i))'\}$ to learn $\mathbf{c}_0$,$\mathbf{c}_c$ and $\mathbf{c}_s$.

Figure 3 shows reconstructions in which the orientation, θ, is varied in Equation 2.

Given a new example with parameters $\mathbf{c}$, we can estimate its orientation as follows. Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $(\mathbf{c}_c|\mathbf{c}_s)$ (thus $\mathbf{R}_c^{-1}(\mathbf{c}_c|\mathbf{c}_s) = \mathbf{I}_2$).

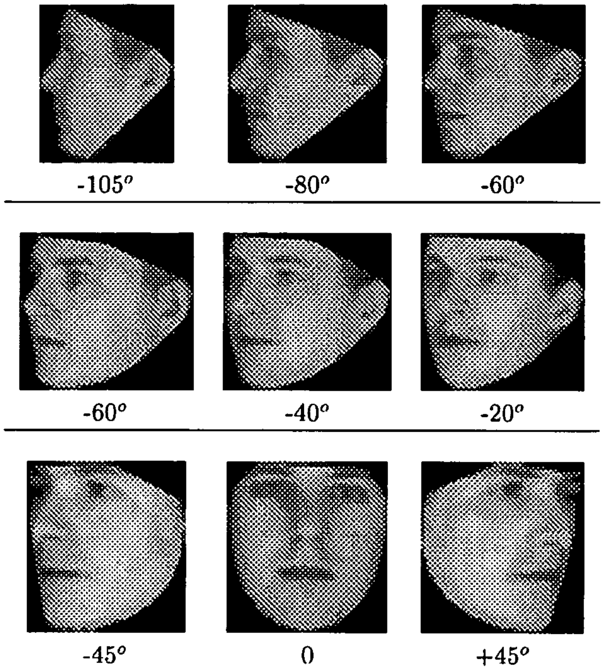


Figure 3. Rotation modes of three face models

Let

$$(x_a, y_a)' = \mathbf{R}_c^{-1}(\mathbf{c} - \mathbf{c}_0) \qquad (3)$$

then the best estimate of the orientation is $\tan^{-1}(y_a/x_a)$.

Figure 4 shows the predicted orientations vs the actual orientations for the training sets for each of the models. It demonstrates that equation 2 is an acceptable model of parameter variation under rotation.

5 Tracking through wide angles

We can use the set of models to track faces through wide angle changes (full left profile to full right profile). We use a simple scheme in which we keep an estimate of the current head orientation and use it to choose which model should be used to match to the next image.

To track a face through a sequence we locate it in the first frame using a global search scheme similar to that described in [3]. This involves placing a model instance centred on each point on a grid across the image, then running a few iterations of the AAM algorithm. Poor fits are discarded and good ones retained for more iterations. This is repeated for each model, and the best fitting model is used to estimate the position and orientation of the head.

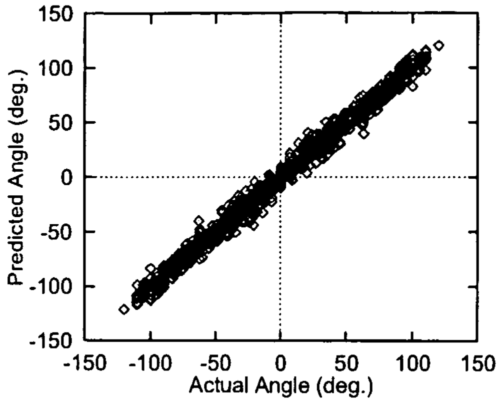


Figure 4. Prediction vs actual angle across training set

Model	Angle Range
Left Profile	-110° - -60°
Left Half-Profile	-60° - -40°
Frontal	-40° - 40°
Right Half-Profile	40° - 60°
Right Profile	60° - 110°

Table 1. Valid angle ranges for each model

We then project the current best model instance into the next frame and run a multi-resolution seach with the AAM. We estimate the head orientation from the results of the search, as described above. We then use the orientation to choose the most appropriate model with which to continue. Each model is valid over a particular range of angles, determined from its training set (see Table 1). If the orientation suggests changing to a new model, we estimate the parameters of the new model from those of the current best fit. We then perform an AAM search to match the new model more accurately. This process is repeated for each subsequent frame, switching to new models as the angle estimate dictates.

When switching to a new model we must estimate the image pose (position, within image orientation and scale) and model parameters of the new example from those of the old. We assume linear relationships which can be determined from the training sets for each model, as long as there are some images (with intermediate head orientations) which belong to the training sets for both models.

Figure 7 shows the results of using the models to track the face in a new test sequence (in this case a previously unseen sequence of a person who is in the training set). The model reconstruction is shown superimposed on frames from the sequence. The methods appears to track well, and is able to reconstruct a convincing simulation of the sequence.

We used this system to track 15 new sequences of the people in the training set. Each sequence contained between 20 and 30 frames. Figure 5 shows the estimate of the angle from tracking against the actual angle. In all but one case the tracking succeeded, and a good estimate of the angle is obtained. In one case the models lost track and were unable to recover.

The system currently works off-line, loading sequences from disk. On a 450MHz Pentium III it runs at about 3 frames per second, though so far little work has been done to optimise this.

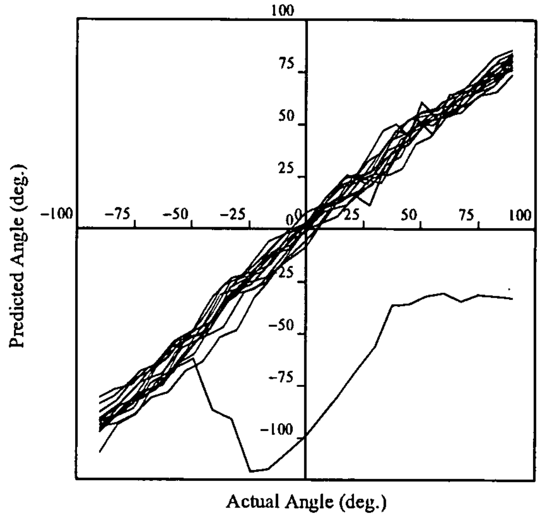


Figure 5. Comparison of angle derived from AAM tracking with actual angle (15 sequences)

6 Predicting Unseen Views

Given a single view of a new person, we can find the best model match and determine their head orientation. We can then use the best model to synthesize new views at any orientation that can be represented by the model. If the best matching parameters are $\mathbf{c}$, we use equation 3 to estimate the angle, θ. Let $\mathbf{c}_{res}$ be the residual vector not explained by the rotation model, ie

$$\mathbf{c}_{res} = \mathbf{c} - (\mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta)) \qquad (4)$$

To reconstruct at a new angle, α, we simply use the parameters

$$\mathbf{c}(\alpha) = \mathbf{c}_0 + \mathbf{c}_c \cos(\alpha) + \mathbf{c}_s \sin(\alpha) + \mathbf{c}_{res} \quad (5)$$

This only allows us to vary the angle in the range defined by the closest model. Since the models all represent the same 3D structure, we anticipate that there will be correlations between parameters for different views of the same individual. To do this effectively we must first project out the effects of pose, lighting etc. A principled approach to this is described in [2]. However, for our experiments, since there is little lighting or expression change in the training set, it is sufficient just to remove the orientation components.

In order to learn the relationship between parameters in one model and those in another, we perform the following steps. For each frame in the training set we use equation 4 to determine the orientation independent component of the parameters for each model. We then compute the mean of such residuals for each person. Let $\hat{\mathbf{c}}_{i,j}$ be the mean of such residuals in the i^{th} model for the j^{th} person. By applying PCA to the means for a given model, we can find the projection, $\mathbf{P}_j$, into an 'identity' sub-space.

Let the projection of each mean in the subspace be

$$\mathbf{b}_{ij} = \mathbf{P}_j^T(\hat{\mathbf{c}}_{i,j} - \hat{\mathbf{c}}_j) \quad (6)$$

where $\hat{\mathbf{c}}_j$ is the mean of the means.

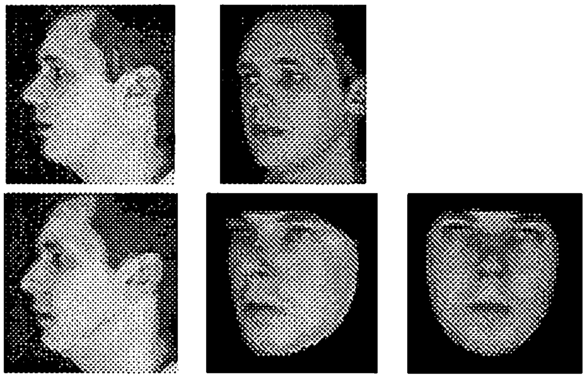
We can use linear regression to learn the relationship which maps each $\mathbf{b}_{ij}$ in the identity space of the j^{th} model to the corresponding mean $\mathbf{b}_{ik}$ in the identity space of the k^{th} model,

$$\mathbf{b}_{ij} = \mathbf{r}_{jk} + \mathbf{R}_{jk}\mathbf{b}_{ik} \quad (7)$$

Thus to reconstruct a new view of a person given a match in a different view;

1. remove the effects of orientation (Eq.4),
2. project into the identity sub-space for the model (Eq.6),
3. project across into the subspace of the target model (Eq.7),
4. project that into the residual space (inverting Eq.6)
5. add the appropriate orientation (Eq. 5).

Figure 6 demonstrates this. Models were built on the data for all but one person. The profile model was then matched to a profile image of the missing person (the reconstruction is shown). The method described above is then used to predict the appearance using the frontal model at two different angles. For comparison, corresponding images of the person at similar angles are shown. Given the small nature of the training set (in this case only 14 people, yielding a 13-D identity space), the results are encouraging.



Best Fit New View New View

Figure 6. The best fit with a profile model is projected to the frontal model to predict new views

7 Discussion and Conclusions

We have demonstrated that a small number of view-based statistical models of appearance can represent the face from a wide range of viewing angles. Although we have concentrated on rotation about a vertical axis, rotation about a horizontal axis (nodding) could easily be included (and probably wouldn't require any extra models for modest rotations). We have shown that the models can be used to track faces through wide angle changes, and that they can be used to predict appearance from new viewpoints given a single image of a person.

So far we have only tested the methods on a relatively small and clean data set. We intend to gather more data in order to obtain better generalisation ability, to include expression and lighting changes and to investigate its performance on more cluttered backgrounds. We hope to obtain better calibrated training images in order to obtain more accurate angle estimates.

We anticipate the approach will be useful in many applications, including driving animated avatars, calculating head pose and making face recognition systems more invariant to viewing angle.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H.Burkhardt and B. Neumann, editors, 5th *European Conference on Computer Vision*, volume 2, pages 484 498. Springer, Berlin, 1998.

[2] N. Costen, T. F. Cootes, G. J. Edwards, and C. J. Taylor. Automatic extraction of the face identity sub-

space. In T. Pridmore and D. Elliman, editors, 10th *British Machine Vison Conference*, volume 1, pages 513–522, Nottingham, UK, Sept. 1999. BMVA Press.

[3] G. Edwards, T. F. Cootes, and C. J. Taylor. Advances in active appearance models. In 7th *International Conference on Computer Vision*, pages 137 142, 1999.

[4] G. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In 3rd *International Conference on Automatic Face and Gesture Recognition 1998*, pages 300–305, Japan, 1998.

[5] D. Graham and N. Allinson. Face recognition from unfamiliar views: Subspace methods and pose dependency. In 3rd *International Conference on Automatic Face and Gesture Recognition 1998*, pages 348 353, Japan, 1998.

[6] H.Murase and S. Nayar. Learning and recognition of 3d objects from appearance. *International Journal of Computer Vision*, pages 5–25, Jan. 1995.

[7] M. J. Jones and T. Poggio. Multidimensional morphable models : A framework for representing and matching object classes. *International Journal of Computer Vision*, 2(29):107 131, 1998.

[8] J. Kwong and S. Gong. Learning support vector machines for a multi-view face model. In T. Pridmore and D. Elliman, editors, 10th *British Machine Vison Conference*, volume 2, pages 503 512, Nottingham, UK, Sept. 1999. BMVA Press.

[9] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic interpretation and coding of face images using flexible models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):743–756, 1997.

[10] T. Maurer and C. von der Malsburg. Tracking and learning graphs and pose on image sequences of faces. In 2nd *International Conference on Automatic Face and Gesture Recognition 1997*, pages 176 181, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.

[11] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. In *SPIE*, volume 2277, pages 12–21, 1994.

[12] S. Romdhani, S. Gong, and A. Psarrou. A multi-view non-linear active shape model using kernel pca. In T. Pridmore and D. Elliman, editors, 10th *British Machine Vison Conference*, volume 2, pages 483 492, Nottingham, UK, Sept. 1999. BMVA Press.

[13] J. Sherrah, S. Gong, and E. Ong. Understanding pose discrimination in similarity space. In T. Pridmore and D. Elliman, editors, 10th *British Machine Vison Conference*, volume 2, pages 523–532, Nottingham, UK, Sept. 1999. BMVA Press.

[14] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71 86, 1991.

[15] T. Vetter. Learning novel views to a single face image. In 2nd *International Conference on Automatic Face and Gesture Recognition 1997*, pages 22–27, Los Alamitos, California, Oct. 1996. IEEE Computer Society Press.

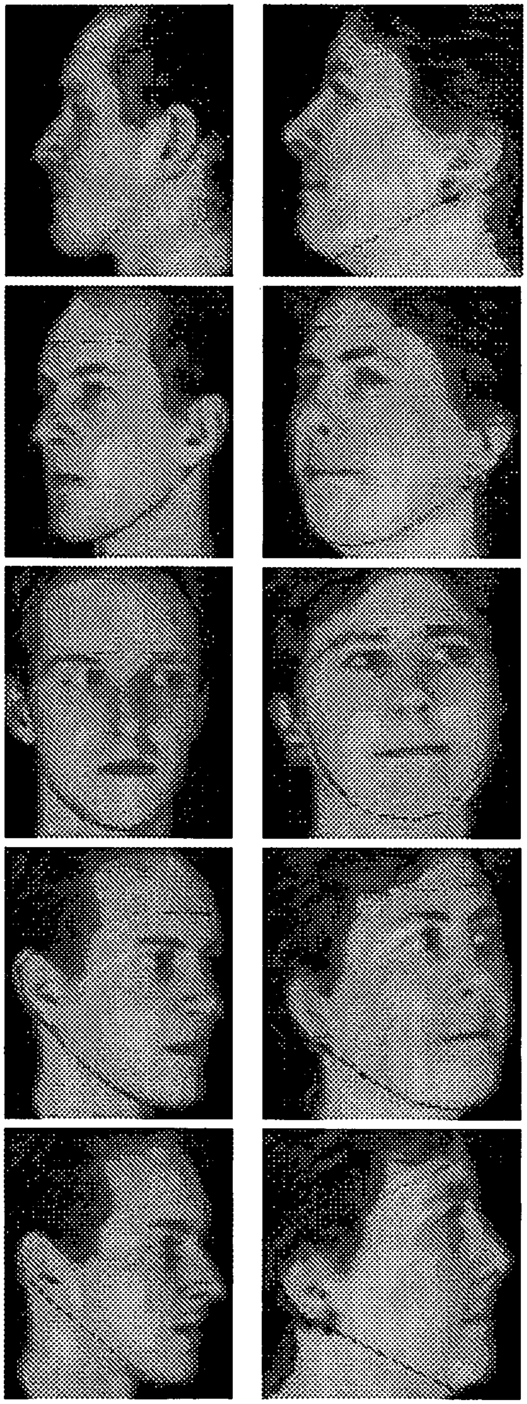


Figure 7. Reconstruction of tracked faces superimposed on sequences

Welcome to IEEE Xplore®

○- Home
○- What Can I Access?
○- Log-out

Tables of Contents

○- Journals & Magazines
○- Conference Proceedings
○- Standards

Search

○- By Author
○- Basic
○- Advanced

Member Services

○- Join IEEE
○- Establish IEEE Web Account
○- Access the IEEE Member Digital Library

Search Results [PDF FULL-TEXT 440 KB] NEXT DOWNLOAD CITATION

Request Permissions
RIGHTSLINK

Face pose estimating system based on eigenspace a

Saito, H. Watanabe, A. Ozawa, S.
Dept. of Inf. & Comput. Sci., Keio Univ., Yokohama, Japan;

This paper appears in: **Image Processing, 1999. ICIP 99. Proceedings. 1 International Conference on**

Meeting Date: 10/24/1999 - 10/28/1999
Publication Date: 1999
Location: Kobe Japan
On page(s): 638 - 642 vol.1
Volume: 1
Reference Cited: 6
Number of Pages: 4 vol.(lxxix+676+977+952+449)
Inspec Accession Number: 6511870

Abstract:
In this paper, we propose a new system for estimating **face pose** from a **faci** In this system, an input **facial image** is compared with a database of **image** **face pose**, then the matched **image** provides the **face pose**. The database c includes not only various **face poses** but also various **illumination** condition the **face pose** estimation system can be used under various **illumination** co For collecting such various **facial images**, they are generated by computer, r using real **images**. The eigenspace method is used for searching a matched i an input **facial image**. Since various **illumination images** are collected in tl of **facial images**, the extracted principle eigenvectors mostly depend on the By performing the matching process in the eigenspace, a matched **image** wit **facial image** can be found. The **pose** of the matched **image** is closest to the The matching process is also fast because it is performed in small dimensiona spanned by selected eigenvectors only. The proposed **pose** estimation systen continuously track the **face pose** of different persons under different **light** co

Index Terms:
eigenvalues and eigenfunctions gesture recognition image matching visual databas eigenspace analysis eigenvectors face pose estimating system facial image illu conditions image database image matching

Documents that cite this document
There are no citing documents available in IEEE Xplore at this time.

Face Pose Estimating System Based on Eigen Space Analysis

Hideo Saito, Akihiro Watanabe, Shinji Ozawa
Department of Information and Computer Science
Keio University
3-14-1 Hiyoshi Kouhoku-ku Yokohama 223-8522, Japan

Abstract

In this paper, we propose a new system for estimating face pose from a facial image. In this system, input facial image is compared with database of images of various face pose, then the matched image provides the face pose. The database of images includes not only various face pose but also various illumination conditions, so that the face pose estimating system can be used under various illumination condition. For collecting such various facial images, they are generated by computer, rather than taking real images. Eigen space method is used for searching the matched image with input facial image. Since various illumination images are collected in the database of facial images, the extracted principle eigen vectors mostly depend on the face pose. By performing the matching process in the eigen space, matched image with the input facial image can be found. The pose of the matched image is most closest to the input face. The matching process is also fast because it is performed in small dimensional space spanned by only selected eigen vectors. The proposed pose estimating system can continuously track the face pose of different person under different light condition.

1 Introduction

Recently, human-computer interface is intensively studied for making computers usable for every people. The recent computers have not only the displays but also cameras for taking images of users. This indicates that such cameras on computer can be used for input device of the user's behavior, so that more natural interface can be realized.

For recognition of user's behavior from the images, automatic face pose estimation technique is one of application of computer vision and image understanding research field. Conventionally, there are many methods for estimating the pose of face from images, which are categorized into : 1.) methods based on detecting of face features such as eyes, noses, mouth, etc., [1][2], 2.) methods based on the intensity distribution of images [5][6]. The former methods generally involve 3D position estimation of the features, thus accurate camera parameters must be known for face pose estimation. Additionally, the feature detection by the image understanding techniques is still hard problem under arbitrary conditions, such as illumination and background scenes. The latter methods are basically model-based method in which the image models obtained previously are used for face pose estimation. Those methods do not need feature detection, but it takes much labors for collecting the models before estimation of the face pose.

In this paper, we propose a new face pose estimating system of model-based method. In this system, computer generated facial images are used for reducing the cost for collecting the images of various models. By the use of computer generated model, it is very easy to change the conditions of the facial images such as illumination. The database of facial images includes not only various face pose but also various illumination conditions, so that the face pose estimating system can be used under various illumination condition. The parametric eigen space method [3] is used for extracting some principle vectors of facial image that mostly depends on the face pose. By performing the matching process in the space spanned by the principle vectors, the input facial image can be matched with images in the database, of which the pose is most closest to the input face.

In this paper, we also show the results of pose estimation from various conditions facial images for demonstrating the efficacy of the proposed system.

2 Face Pose Estimation Based on Parametric Eigen Vector Method

Murase et.al. proposed the parametric eigen space technique [3] in which important feature vectors (principle vectors) of the database can be extracted by the eigen space analysis. They apply this method to object recognition from the appearance in which each model identity is represented in the eigen space spanned by the principle vectors. In this method, the storage capacity of the database can be reduced because the extracted important feature vectors in small dimension are only required for matching procedure of the recognition. This reduction also helps to reduce the computation cost for matching of input features with the features of database. In this method, KL transform is employed for extract several important principle vectors of the data set in the database. Consequently, the objects can be represented with small dimensions (i.e. 20) rather than the original dimensions, (i.e. 256^2).

In our face pose estimation system, the parametric eigen space method is used for extracting principle vectors such that the principle vectors face mostly depends on the face pose.

0-7803-5467-2/99/ $ 10.00 © 1999 IEEE

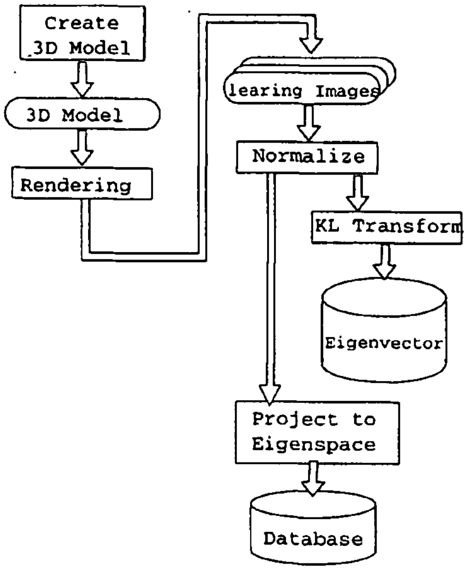


Figure 1: The flowchart of the learning stage.

An intensity distribution of image mostly depends on the face pose, personality of face, and illumination condition. The personality means that personal features such as the position of the mouse, eyes, and nose, the shape of the face, etc. However, such personality is difficult to be represented by the small dimension principle vector because of the complication of the personal features. Thus, such personal features is difficult to be extracted as principle feature vectors by the eigen space technique.

The illumination condition mainly affects to the intensity distribution of image. However, if the set of facial images include images under variety illumination conditions, the contribution of the illumination condition to the principle feature vectors can be reduced.

In this way, face pose is estimated by constructing the eigen space spanned by the principle vectors extracted from facial image data taken under various illumination.

3 Face Pose Estimating System

The proposed pose estimation system is divided into two stages; learning stage, and estimation stage. In the learning stage, the principle vectors are extracted from the facial image data base. In the estimation stage, the face pose of the input image is estimated in the eigen space that is spanned by the principle vectors.

Figure 1 and figure 2 show the flow of each stage.

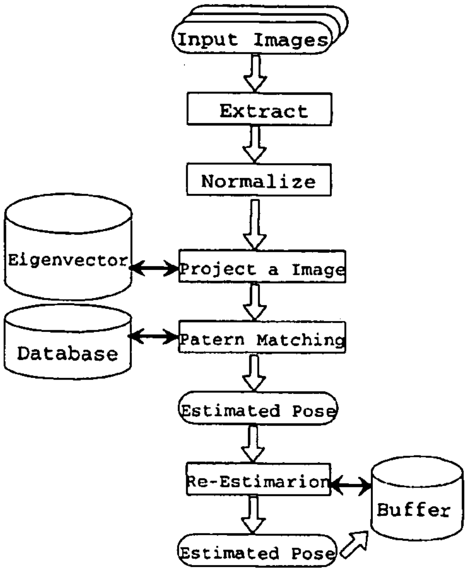


Figure 2: The flowchart of the estimation stage.

3.1 Learning stage

3.1.1 Collection of facial images

The facial images under various illumination and various face pose are collected by computer graphics (CG). The use of the computer generated facial images for the database has two advantages:

- Collection of various facial images is easy.
- Control of face pose and environmental condition is easy.

If we need to collect various facial images from real human, much labor effort is required for collection of the images. We need to know the angle of face pose for each facial image in the database used for the learning. The angle of face pose in CG image is easily to be obtained, while the angle in real facial image is difficult. We also need to collect various face pose images under various environment such as illumination. It is easy to change such environment for the computer generated facial images. For such purposes, we generate the facial images by computer graphics rather than taking real images.

Although various face shapes can reduce the dependency of the feature of face shape to the principle feature vectors, we use only one face shape model in this paper. This is because that we assume that difference in different person's face shape gives much smaller effect to the difference in intensity distribution than the face pose difference.

3.1.2 Extraction of eigen space

An facial image $\mathbf{x}$ with $n \times n$ pixels is defined as

$$\mathbf{x} = [x_1, x_2, \cdots, x_{n^2}]. \qquad (1)$$

The set of N facial images is indicated as

$$[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]. \qquad (2)$$

The average value $\mathbf{c}$ in the images are subtracted from every image, then the matrix $\mathbf{X}$ with n^2 rows $\times N$ columns is represented as

$$\mathbf{X} = [\mathbf{x}_1 - \mathbf{c}, \mathbf{x}_2 - \mathbf{c}, \cdots, \mathbf{x}_N - \mathbf{c}]. \qquad (3)$$

The covariance matrix $\mathbf{Q}$ of $\mathbf{X}$ is derived as

$$\mathbf{Q} = \mathbf{X}\mathbf{X}^T. \qquad (4)$$

Then the eigen equation can be shown as

$$\lambda_i \mathbf{e}_i = \mathbf{Q}\mathbf{e}_i. \qquad (5)$$

With this equation, eigen values and eigen vectors can be calculated so that eigen space with dimension k ($<< n^2$) can be constructed.

After the construction of eigen space, the images are projected onto the eigen space for making the database of the eigenvectors of all the facial images. Because the dimension k of the eigen space is much smaller than the dimension of the image n^2, the required storage for the database can be reduced by the factor of k/n^2.

3.2 Pose Estimation Stage

3.2.1 Extraction and normalization of facial image

For estimating of face pose, face area must be extracted from input image. The face area is defined from brow to chin and from left ear to right ear.

For extracting the face area from input image, color information of the input image is used. First, the face candidate region is extracted by thresholding hue and saturation of the input image. The threshold values for hue and saturation are defined according to the distribution of face color. The extracted regions, which is represented in binary mask image, include not only the face region but also some regions of non facial objects. To remove the non facial regions, size and shape of every region is calculated after the labeling of the regions. The face area is selected based on the sise and shape of the region.

3.2.2 Pose Estimation

The normalized input image $\mathbf{y}$ is projected onto the eigen space obtained as the previous section.

$$\mathbf{z} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T (\mathbf{y} - \mathbf{c}). \qquad (6)$$

For estimating the pose of the input image, the Euclid distance between eigen vector of input image $\mathbf{z}$ and that of images in the database $\mathbf{g}_i$.

$$d_i = |\mathbf{z} - \mathbf{g}_i|. \qquad (7)$$

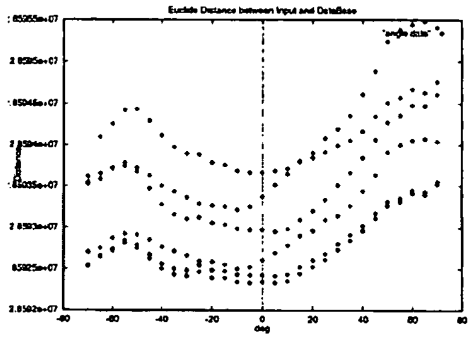


Figure 3: Euclid distance between input image and database images in Eigenspace.

The pose of the facial image which has minimum distance d_i is initial guess of the face pose.

This matching calculation cost can be reduced because the distance is calculated in the eigen space of dimension k.

In figure 3, an example of the Euclid distance between input image and the images of the database in eigen space is shown. The face angle at smallest distance is the initial guess of face pose.

The initial guess of the face pose is compared with the estimated face pose at the previous image frame. If the pose difference is larger than pre-determined threshold value, a local minimum is searched around the face pose angle of the previous image frame, and the angle of the local minimum is the final estimation of the face pose. Such correction based on the temporal continuity of the face pose avoids eventual error in face pose estimation. The flow of this procedure is shown in figure 4.

4 Experiments

4.1 Experimental Conditions

In this experiment, we prepare database of images of 29 different face angles (-70 deg.$\leq \theta \leq$ 70 deg. , at 5 deg. interval) under 6 different illumination conditions, then 174 facial images in total. Figure 5 shows example images of the facial images. The illumination condition is changed by the combination of point light source and ambient light source at different position as shown in figure 6.

Since various illumination images are collected in the set of facial images, the extracted principle eigen vectors mostly depend on the face pose. Therefore, the proposed system can be used under arbitrary illumination condition.

For determining the proper dimension of eigen space, we performed test experiment. In the test experiment, we investigated the relationship between the estimated pose angle and dimension of the eigen space. Figure 7 shows the result of the test experiment. This result shows that there is no difference in the case of

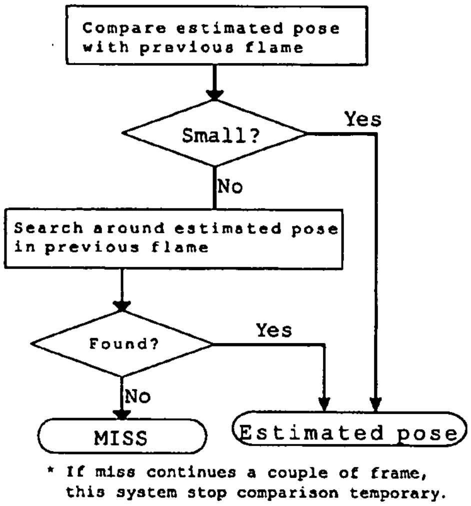

Compare estimated pose
with pravious flame
Small?
Yes
No
Search around estimated pose
in previous flame
Found?
Yes
No
MISS
Estimated pose
* If miss continues a couple of frame,
this system stop comparison temporary.

Figure 4: Correction of pose estimation.

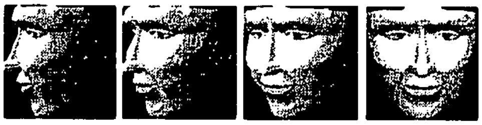


Figure 5: Examples of facial images in the database.

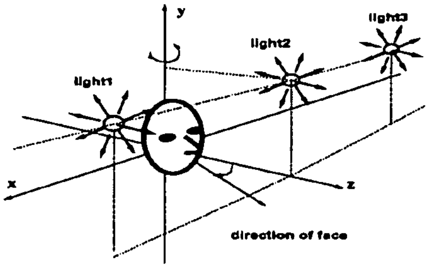


Figure 6: Position of the light source for generating.

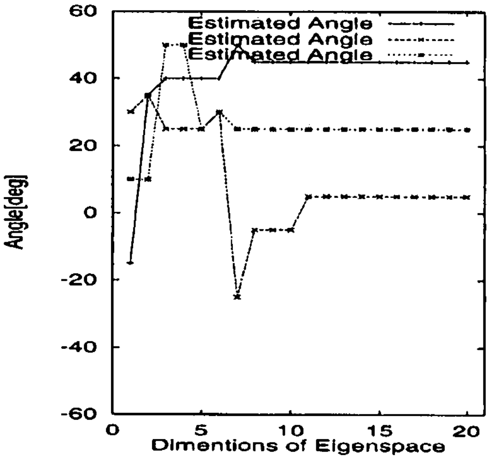


Figure 7: Relationship between the dimension of the eigen space and estimated face pose

dimension more than 11. From this result, we conclude that 20 is sufficient for estimating face pose in this experiment. Figure 8 shows the basis images under 4th dimension in this eigen space.

4.2 Pose Estimation Results

In figure 9, examples of the system results are shown. Top left image is input image, and top right image is masked image extracted by thresholding hue and saturation. Although the face pose estimation is performed using gray images, the color information of input image is used for automatic extraction of the face region. Bottom left image is input to face pose estimation process, which is normalized into size of 128 × 128 pixels. Bottom right image is matched image in the database. Other examples of pose estimation by the proposed system are also shown in figure 10.

The pose face estimation is continuously performed to input image sequence. The estimation is sometimes wrong, but such wrong estimation can be corrected by checking the difference with the previous estimation. By the correction of the wrong estimation, the tracking of the face pose can be performed in reasonable quality.

In figures 9 and 10, examples for the face with grasses are shown. Those cases demonstrate that the detail feature does not affect to the pose estimation because the eigen space method can only extract the feature depending on the face pose. Those examples also show the robustness in the face pose estimation in our system.

This system is actually constructed in SGI-O2 (R5000, 180MHz) with O2 Cam. It takes about 2 seconds for face pose estimation of 1 frame.

5 Conclusion

We propose a method for estimation of face pose using eigen space method.

In this method, the computation cost and storage capacity are much smaller than correlative matching method in the image space, because the matching is performed in the eigen space of small dimensions. Furthermore, since the eigen space is mostly depending on only the feature of face pose difference, the detailed

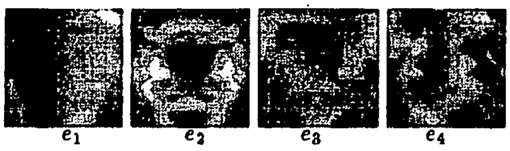


Figure 8: Basis images of eigen space.

deference in the input images, i.e. the existence of grass, do not affect to the pose estimation.

In this method, 3D geometrical information is not required to estimate face pose, because pose estimation is performed according to the appearance of the face in 2D image.

Computer graphics is employed for collecting facial images for learning in this system. The use of computer graphics reduces much effort to collect facial images under various situation. These images are not real images, but it is enough to estimate the face pose.

References

[1] A. Aoyama, T. Yamamura, N. Ohnishi, and N.Sugie: "Gaze Estimation from Single Camera", IEICE Technical Report, PRU95-23, pp.131-136(1996) (In Japanese)

[2] Andrew Gee, Roberto Cipolla: "Determining the gaze of faces in images". Image and Vision Computing Vol.12, No.10, pp.639-647(1994)

[3] Hiroshi Murase, Shree K. Nayar: "Parametric Eigenspace Representation for Visual Learning and Recognition". SPIE Vol.2031 Geometric Method in Computer Vision II, pp.378-391 (1993)

[4] Hiroshi Murase, Shree K. Nayar: "Illumination Planning for Object Recognition in Structured Environments". IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.31-38 (1994)

[5] Kaori Suzuki, Hideo Saito, Shinji Ozawa: "Estimation of face orientation from shading images", Transactions of the Institute of Electrical Engineers of Japan, Part C, vol.117-C, no.10, pp. 1377-1383 (1997) (In Japanese)

[6] Akitoshi Tsukamoto, Chil-Woo Lee, Saburo Tsuji: "Detection and pose estimation of human face with synthesized image models", Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol.1, pp.754-757(1994)

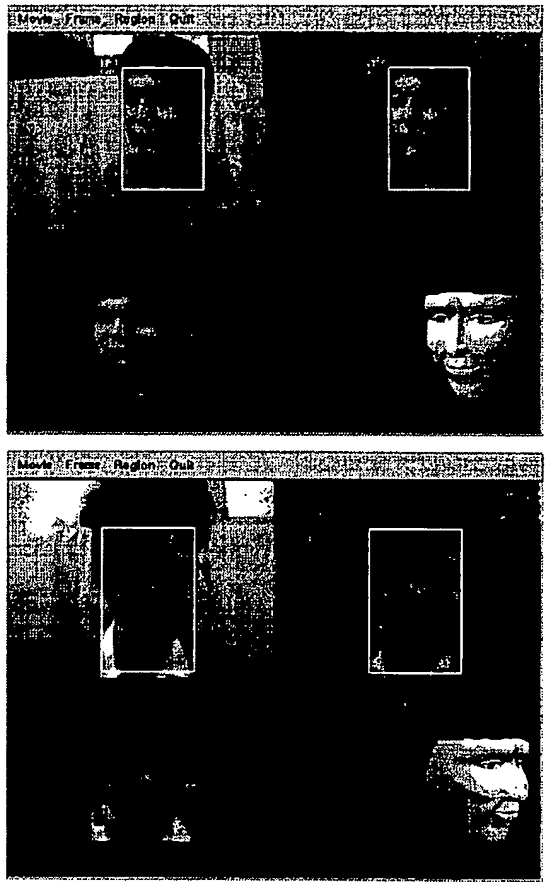


Figure 9: Input image (top left), extracted face area (top right), normalized facial image (bottom left) and estimated pose (bottom right). The rectangle area represents the extracted area as a face region.

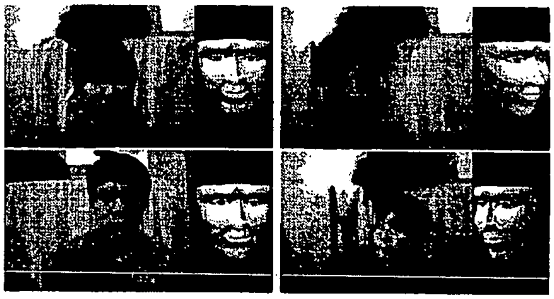


Figure 10: Results of pose estimation.

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

□ BLACK BORDERS

□ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

☑ FADED TEXT OR DRAWING

□ BLURRED OR ILLEGIBLE TEXT OR DRAWING

□ SKEWED/SLANTED IMAGES

□ COLOR OR BLACK AND WHITE PHOTOGRAPHS

□ GRAY SCALE DOCUMENTS

□ LINES OR MARKS ON ORIGINAL DOCUMENT

□ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

□ OTHER: ____________________

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.